Metric Learning With Siamese Networks for Re-Identification and Tracking

by

Madhu KIRAN

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, 12 SEPTEMBER 2024

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Eric Granger, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Rafael Menelau Oliveira e Cruz, Thesis Co-Supervisor
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Stéphane Coulombe, Chair, Board of Examiners
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Marco Pedersoli, Member of the Jury
Department of Systems Engineering, École de technologie supérieure

Mr. Jim Clark, External Examiner
Department of Electrical and Computer Engineering, McGill University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON 13 JUNE 2024

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to my esteemed supervisors, Prof. Eric Granger and Prof. Rafael Menelau Cruz, for their unwavering support and encouragement. Their invaluable guidance has played a crucial role in the success of my work, particularly with regard to research methods and providing direction throughout my thesis. I appreciate their commitment, as demonstrated through numerous discussions and collaborative efforts resulting in multiple published research articles.

I would also like to extend special thanks to my friend and lab mate, Praveen, whose assistance proved crucial during my Ph.D. journey.

My most profound appreciation goes to my parents for their boundless love, sacrifices, and unwavering support. I am incredibly thankful to my wife, Sudha, whose understanding and sacrifices enabled me to dedicate long hours in the lab and pursue this research. Completing this thesis would have been challenging without her support. I also express my gratitude to my daughter, Yalini, for the joy and strength she brought, making the challenging moments of this thesis more bearable.

Apprentissage métrique avec des réseaux siamois pour la ré-identification et le suivi

Madhu KIRAN

RÉSUMÉ

Le suivi d'objets visuels (VOT) et la réidentification de personnes (ReID) sont essentiels pour une large gamme de surveillances et de suivis vidéo, tels que le suivi de cibles multi-caméras, le suivi de piétons pour la conduite autonome, le contrôle d'accès biométrique, etc. Une fois une personne détectée, le VOT produit des cadres englobants continus pour cette personne sur une séquence vidéo à partir d'un emplacement initial. Parallèlement, le ReID identifie les traces perdues d'une personne ou reconnaît des personnes ou objets observés à travers différentes caméras. Ils rencontrent des défis communs, notamment dans les scénarios avec occlusion et changements d'apparence de la cible. Les objectifs d'apprentissage pour le suivi et le ReID sont similaires aux tâches de correspondance de similarités utilisant des architectures de réseaux Siamese. La mise à jour d'un traceur est effectuée en ligne car l'apparence de l'objet ou le prototype doit être appris au fil du temps avec les changements d'apparence. Le ReID n'implique pas d'apprentissage en ligne explicite; il s'agit plutôt d'utiliser des séquences vidéo pour représenter au mieux un objet et le faire correspondre à une base de données d'objets.

Les défis liés à l'apprentissage en ligne dans le suivi, où des méthodes comme les approches classiques ou les méthodes de réseau neuronal convolutif profond (CNN) visent à apprendre l'apparence de la cible pendant le suivi pour éviter la dérive de la cible au fil du temps. Cependant, l'adaptation du modèle à l'aide d'échantillons d'un traceur peut être entravée par le bruit introduit par la dérive. Détecter et gérer les dérives devient crucial, et une sélection inappropriée des échantillons peut affecter la robustesse du modèle. Par exemple, une diversité inadéquate dans les échantillons de formation lors de l'adaptation peut entraîner une dérive significative du modèle lorsque l'apparence de la cible change en raison de variations, par exemple, de l'éclairage et du point de vue.

Dans la ReID de personnes basée sur la vidéo, l'utilisation de tracklets de personnes dans les requêtes peut relever des défis tels que l'occlusion, le positionnement inexact des cadres englobants, et les variations d'apparence dues aux changements d'éclairage et de point de vue. Des améliorations sont recherchées dans les méthodes d'agrégation de caractéristiques pour les séquences vidéo, en tenant compte des limites de la longueur des séquences et du risque de surajustement. L'occlusion pose un défi important dans l'apprentissage en ligne pour le suivi et le ReID de personnes, amenant les extracteurs de caractéristiques à se concentrer sur des régions non-objectives, conduisant à des correspondances erronées. Développer des solutions efficaces du point de vue computationnel pour gérer l'occlusion est crucial pour les applications en temps réel. Cette thèse se concentre sur trois problèmes principaux : l'apprentissage en ligne avec sélection dynamique de modèles/échantillons pour le suivi, la représentation vidéo et l'apprentissage de représentations conscientes de l'occlusion.

La première contribution de cette thèse se concentre principalement sur l'amélioration de l'apprentissage en ligne dans les modèles VOT, avec un accent particulier sur les défis liés à

la dérive conceptuelle et à l'occlusion. La dérive conceptuelle englobe les changements dans l'apparence de l'objet suivi, catégorisés comme graduels, abrupts et récurrents. Notre recherche souligne les avantages de l'adaptabilité de la dérive conceptuelle graduelle tout en reconnaissant que les changements abrupts résultent souvent de distractions comme l'occlusion, nécessitant une gestion prudente des mises à jour du modèle. De plus, les dérives récurrentes se produisent lorsque l'apparence précédente de l'objet se reproduit, et nous avons constaté qu'il est efficace de maintenir un tampon d'échantillons à haute variance pour le suivi en ligne.

La deuxième contribution de cette thèse se concentre sur la ReID de personnes dans l'analyse vidéo et la surveillance, visant à surmonter les limitations existantes telles que l'apparence changeante, la variation de point de vue à travers les caméras, et l'occlusion. Pour y remédier, incorporez les motifs de mouvement des individus comme indice supplémentaire pour le ReID. Notre solution proposée introduit le réseau d'Attention Mutuelle Guidée par le Flux, fusionnant les séquences de cadres englobants et de flux optique sur des tracklets. Ceci utilise un backbone CNN 2D pour coder à la fois les informations d'apparence temporelle et spatiale. De plus, nous présentons une méthode novatrice pour agréger les caractéristiques à partir de flux d'entrée étendus pour améliorer la représentation au niveau de la séquence vidéo. Les résultats expérimentaux montrent une amélioration significative de la précision du ReID par rapport aux réseaux d'attention traditionnels et aux méthodes actuelles de pointe dans le ReID de personnes basé sur la vidéo. Cette recherche met en évidence le potentiel des mécanismes d'attention guidés par des motifs de mouvement pour améliorer les capacités des modèles d'apprentissage profond pour des applications robustes de ReID vidéo.

Notre prochaine introduction concerne un modèle étudiant-enseignant Holistique-Génératif (HG) conçu pour le ReID de personnes occlues, éliminant le besoin d'étiquettes d'identité d'image et de processus intensifs en ressources axés uniquement sur les parties visibles des régions occlues. L'enseignant HG proposé utilise la Distribution de Classe Discriminative (DCD) à travers des échantillons dans un ensemble de données complet pour former un modèle étudiant, permettant la génération de cartes d'attention et abordant les défis posés par l'occlusion. Contrairement aux approches prévalentes dans la littérature qui utilisent une supervision externe comme la pose pour des indices de visibilité, notre méthode repose sur la distribution holistique des données pendant l'entraînement, la traitant comme une étiquette douce. Par conséquent, lors des tests, notre modèle fonctionne sans problème sans avoir besoin d'indices externes comme la pose, avec les paramètres globaux comprenant uniquement l'encodeur de base et un embedding compact pour la génération de cartes d'attention lors de l'extraction de caractéristiques. De plus, l'apprentissage conjoint d'un autoencodeur de débruitage améliore la capacité du modèle à se régénérer de l'occlusion. Les évaluations empiriques sur des ensembles de données divers et difficiles démontrent la performance supérieure de notre méthodologie HG, surpassant les modèles de pointe dans les tâches Occluded-ReID et Holistic ReID.

Notre dernière contribution explore l'espace pour la dissimilarité dans le ReID de personnes afin de résoudre le problème de chevauchement entre les classes causé par des modèles efficaces sur le plan computationnel avec des capacités relativement plus petites et des données d'entraînement limitées. Nous proposons d'appliquer une transformation de dichotomie à l'espace des

caractéristiques et de classer les paires d'échantillons comme similaires ou dissimilaires avec un classificateur à marge maximale. L'apprentissage de bout en bout d'un classificateur dans l'espace de la dissimilarité a été observé pour améliorer la précision de récupération pour les tâches de ReID de personnes.

Les résultats de cette thèse indiquent qu'un apprentissage en ligne efficace avec une sélection soignée des échantillons basée sur des techniques de détection de dérive peut permettre un suivi à long terme avec des mises à jour minimales du traceur, entraînant ainsi une faible complexité globale. De plus, il a également été démontré qu'une bonne représentation peut être apprise en choisissant de longues séquences vidéo. L'occlusion représente un défi dans les applications de suivi et de ReID. Il a été démontré qu'avec un apprentissage conscient de l'occlusion, il peut y avoir une amélioration globale à la fois des performances de suivi et de ReID. Cela résout ce problème pour le scénario pratique avec des données étiquetées minimales d'occlusion.

Mots-clés: Suivi d'objets visuels, réidentification de personnes, apprentissage en ligne, apprentissage conscient de l'occlusion, réseaux siamois, flux optique, espace de dissimilarité.

Metric Learning With Siamese Networks for Re-Identification and Tracking

Madhu KIRAN

ABSTRACT

Visual object tracking (VOT) and person re-identification (ReID) go hand in hand for a wide range of monitoring and video surveillance, like multi-camera target tracking, pedestrian tracking for autonomous driving, access control in biometrics, etc. Once a person has been detected, VOT produces continuous bounding boxes for that person over a video sequence given an initial location. At the same time, ReID identifies lost tracks of a person or identifies persons or objects seen in different cameras. They suffer from common challenges, particularly in scenarios with occlusion and changing target appearances. Learning objectives for tracking and ReID are similar to similarity-matching tasks using Siamese network architectures. The update of a tracker is performed online as object appearance or prototype needs to be learned over time with changes in appearance. ReID does not involve explicit online learning; it is more about using video sequences to represent an object best and match it with a database of objects.

Challenges related to online learning in tracking, where methods like classical approaches or deep convolutional neural network (CNN) methods aim to learn target appearance during tracking to prevent target drift over time. However, model adaptation using samples from a tracker can be hindered by noise introduced through drifting. Detecting and managing drifts becomes crucial, and improper sample selection can impact model robustness. For instance, inadequate diversity in training samples during adaptation can significantly lead to model drift when the target's appearance changes due to variations in, e.g., lighting and viewpoint. In video-based person ReID, leveraging person tracklets in queries can address challenges like occlusion, inaccurate bounding box positioning, and variations in appearance due to lighting and viewpoint changes. Improvements are sought in feature aggregation methods for video sequences, considering limitations in sequence length and potential overfitting. Occlusion poses a significant challenge in online learning for tracking and person ReID, causing feature extractors to focus on non-object regions, leading to false matches. Developing computationally efficient solutions for handling occlusion is crucial for real-time applications. This thesis focused on three main problems: online learning with dynamic template/sample selection for tracking, video representation, and occlusion-aware representation learning.

The first contribution in this thesis primarily concentrates on enhancing online learning in VOT models, with a specific focus on addressing challenges related to concept drift and occlusion. Concept drift encompasses changes in the appearance of the tracked object, categorized as gradual, abrupt, and recurring. Our research highlights the adaptability benefits of gradual concept drift while acknowledging that abrupt changes often result from distractions like occlusion, necessitating cautious handling of model updates. Moreover, recurring drifts occur when the object's previous appearance reoccurs, and we found that maintaining a sample buffer with high variance proves effective for online tracking.

The second contribution in this thesis focuses on video person ReID within video analytics and surveillance, aiming to overcome existing limitations such as changing appearance, viewpoint variation across cameras, and occlusion. To address this, incorporate the motion patterns of individuals as an additional cue for ReID. Our proposed solution introduces the Flow-Guided Mutual Attention network, merging bounding box and optical flow sequences over tracklets. This utilizes a 2D-CNN backbone to encode both temporal and spatial appearance information. Furthermore, we present a novel method for aggregating features from extended input streams to enhance video sequence-level representation. Experimental results show a significant improvement in ReID accuracy compared to traditional gated-attention networks and current state-of-the-art methods in video-based person ReID. This research highlights the potential of motion-pattern-guided attention mechanisms in enhancing the capabilities of deep learning models for robust video ReID applications.

Our next introduces a novel Holistic-Generative (HG) student-teacher model designed for occluded person ReID, eliminating the need for image identity labels and resource-intensive processes focused solely on visible parts of occluded regions. The proposed HG teacher uses the Distribution of CLass Distances (DCD) across samples in a comprehensive dataset to train a student model, allowing the generation of attention maps and addressing challenges posed by occlusion. Unlike prevalent approaches in the literature that use external supervision like pose for visibility cues, our method relies on holistic data distribution during training, treating it as a soft label. Consequently, during testing, our model seamlessly operates without needing external cues like pose, with the overall parameters comprising only the backbone Encoder and a compact embedding for attention map generation during feature extraction. Additionally, joint learning of a denoising autoencoder enhances the model's ability to self-recover from occlusion. Empirical evaluations on diverse and challenging datasets demonstrate the superior performance of our HG methodology, surpassing state-of-the-art models in Occluded-ReID and Holistic ReID tasks.

Our final contribution explores the space for dissimilarity in Person ReID to solve the problem of overlap between classes caused by computationally efficient models with relatively smaller capacities and limited training data. We propose applying dichotomy transformation to feature space and classifying sample pairs as similar or dissimilar with a max-margin classifier. End-to-end learning of a classifier in the dissimilarity space has been observed to improve retrieval accuracy for Person ReID tasks.

Results in this thesis indicate that effective online learning with careful sample selection based on techniques for drift detection can allow long-term tracking with minimal updates to the tracker, thereby low overall complexity. In addition, it has also been shown that a good representation can be learned by choosing long video sequences. Occlusion is a challenge in both tracking and ReID applications. It has been shown that with occlusion-aware learning, there can be an overall improvement in both tracking and ReID performances. It solves this problem for the practical scenario with minimalistic occlusion labeled data.

Keywords: visual object tracking, person re-identification, online learning, occlusion aware learning, siamese networks, optical flow, dissimilarity space

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

VOT	Visual Object Tracking
ReID	Re-Identification
DCD	Distribution of Class Distances
HG	Holistic Guidace
VS	Video Surveillance
ROI	Region Of Interest
RNN	Recurrent Neural Netowork
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
FC	Fully Connected
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
DL	Deep Learning
MMD	Maximum Mean Discrepancy
KL	Kullback-Leibler
FIFO	First In, First Out
KDE	Kernel Density Estimation
AUC	Area Under Curve
CMC	Cumulative Matching Characteristics

mAP	mean Average Precision
DML	Deep Metric Learning
FLOPS	Floating Point Operations Per Second

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

y	Class label
φ	Image embedding for matching or classification
I_t	Input image at time frame t
ψ_q	Intermediate image embedding
$\mathbf{D}^{\text{wc}}$	Distribution of within-class examples
$\mathbf{D}^{\text{bc}}$	Distribution of within-class examples
W	Weights of fully connected layer
$\mathcal{A}$	Auxillary target template buffer
$\mathbf{B}$	Main target template buffer
$\epsilon(t)$	Difference between error measured at time frame t

INTRODUCTION

Due to rapid technological progress, video surveillance (VS) has become integral to ensuring safety in modern security systems. The widespread use of surveillance cameras in public and private spaces has led to unprecedented video data, presenting opportunities and challenges. Video surveillance is pivotal in safeguarding public safety, deterring criminal activities, and aiding investigations. However, the sheer volume of visual information captured with modern video surveillance cameras necessitates sophisticated techniques for efficient processing, analysis, and interpretation.

Central to the efficacy of VS is the ability to accurately and consistently track individuals as they navigate dynamic environments (Kiran *et al.*, 2021a). The result of such tracking would be a set of trajectories. Trajectories are collections of the same person or object's ROIs(Region of interests or cutouts). This demands the seamless integration of two key components: person re-identification (re-ID) and tracking (Bewley, Ge, Ott, Ramos & Upcroft, 2016a). Person re-ID refers to recognizing individuals across different camera views or instances, even under significant variations in lighting, pose in addition to the presence of occlusions Ye *et al.* (2021). Tracking involves continuously monitoring an individual's movement, traversing the surveillance field (Ciaparrone *et al.*, 2020). The symbiotic relationship between person ReID and tracking is a linchpin for achieving reliable and comprehensive surveillance outcomes (Ye *et al.*, 2021).

Fig. 0.1 illustrates a VS system to generate person trajectories. The person detector produces bounding boxes or ROIs of persons on the scene. This initializes the tracker to track the person independent of the detector to produce ROIs across different time frames. The Person Re-Identification module identifies persons from different time frames or different cameras from a surveillance network and produces trajectories. A trajectory contains information on the person's identity and location in the video frame(ROI) for various time frames.

In general, VS systems suffer from various challenges, such as occlusion, changing appearance, changes in lighting, and camera viewpoints. The amalgamation of person-ReID (ReID) and tracking transcends these limitations by enabling the recognition and monitoring of individuals across multiple cameras and over extended periods. Tracking helps follow objects locally in a scene. In contrast, ReID helps re-identify objects lost by the tracker due to occlusion or temporary disappearance from the scene and identifies objects across two cameras. Hence, tracking and ReID complement each other (Shuai, Berneshawi, Modolo & Tighe, 2020; Wang *et al.*, 2021c). Person Re-ID and tracking are strongly related because they are similarity-matching problems, and most of the related works use Siamese networks (Ondrašovič & Tarábek, 2021; Ye *et al.*, 2021). In addition, the challenges faced by tracking and ReID are similar, such as changing the appearance of objects/persons and occlusion and camera viewpoints.

Deep Siamese networks have been applied in a broad range of applications owing to their ability to produce invariant and representative embeddings. Siamese networks can extract features of images with unknown class distribution (Koch *et al.*, 2015) and can be trained on generic images for a specific task without needing domain-specific knowledge. Visual Object tracking with Siamese networks has been a promising approach to tracking diverse objects at above real-time

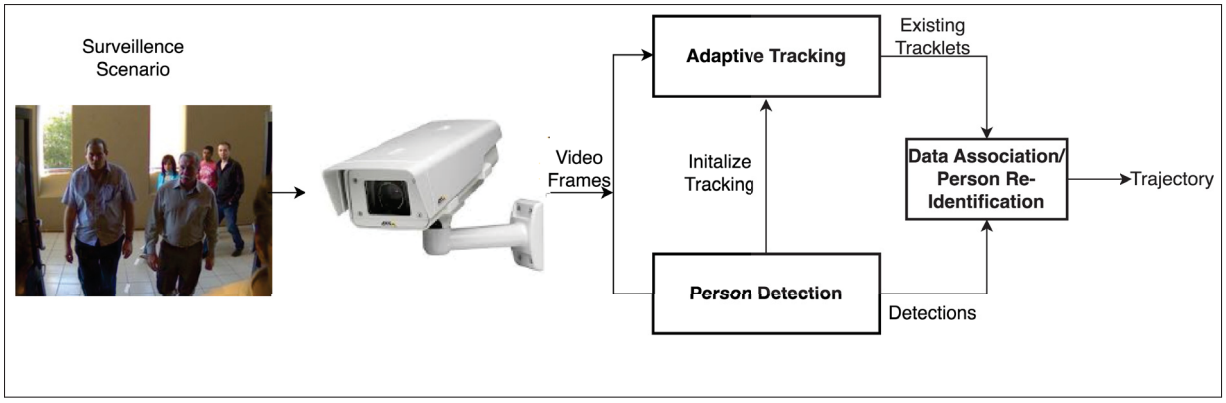

Figure 0.1 Illustration of a detector, tracking, and person-reidentification module in a video surveillance system to construct person trajectories. Such trajectories can later be subjected to analysis for surveillance

speeds (Zhang & Peng, 2019). Similarly, the representation power of Siamese networks has been monumental in matching persons from different cameras, partial occlusion, and other challenging scenarios (Ye *et al.*, 2021). The priority of challenges that need to be addressed for tracking and ReID can differ due to the dependency on each other in a VS application. For example, for tracking, online learning (Bhat, Danelljan, Gool & Timofte, 2019a) is necessary for robust tracking, while for ReID, managing some of the challenges like Occlusion (Zhuo, Chen, Lai & Wang, 2018b) will be a priority.

This thesis explores the two major components of an object/person tracking system for VS, i.e., person re-ID and tracking. By harnessing the representation power of Siamese Networks, this research seeks to enhance surveillance systems' accuracy, efficiency, and reliability by addressing the real-world challenges faced by tracking and ReID.

The subsequent chapters of this thesis delve into the foundational principles of person re-ID and tracking, investigate state-of-the-art methodologies and algorithms, and propose novel approaches to overcome existing limitations. Moreover, real-world applications of these advancements in law enforcement, public safety, retail analytics, and beyond will be explored, demonstrating the far-reaching impact of this research.

0.1 Problem Statement and Challenges

Re-identification (ReID) and tracking in video surveillance are fundamentally concerned with matching image patches, a challenge that can be modeled as a pattern recognition problem. In these tasks, the goal is to analyze pairs of image patches that may represent different perspectives of the same object, person, or distinct entities altogether. The core objective is to develop a method for learning feature representations, denoted as ϕ, in such a way that the representations of the same object or person (viewed from different angles or in different frames) are more similar to each other than they are to representations of different objects or unrelated background

elements. This involves training the system to discern and minimize the distance between feature representations of the same entity while maximizing the distance between those of different entities. ReID and tracking share this paradigm, although they apply it in slightly different contexts: ReID focuses on identifying the same individual across different scenes or time frames, whereas tracking is concerned with continuously following an individual or object within a scene.

For tracking, both image patches are obtained from images taken by the same camera in the given video. Tracking is focused on finding the location of patches that best match the object to obtain the object-bounding box. Fig. 0.2 shows the general siamese tracker. I_z and I_x are template and test image patches. $\phi(z)$ and $\phi(x)$ are the corresponding embeddings. The objective is to match the best test image patch that matches the template embedding by correlating the pairs of embeddings. The test image patches have been sampled around the target's last known location. Since the image patches are obtained from consecutive image frames from a video, the appearance change object between patches is assumed to be gradual. Since candidate image patches for the next time frame are selected based on the predicted bounding box of the current time frame, errors in matching can lead to increasing localization errors(drifting) of candidate image patches, leading to track failure. Candidate feature representations are often learned or adapted over time to accommodate the object's changing appearance and avoid drifting. As the model adapter, this is shown in Fig. 0.2.

In ReID, the patches can be from the same or different cameras; hence, the object/person's appearance change can be large due to changing viewpoints. Fig 0.3 shows a general ReId system. Query image patches or cutouts are represented by $I_q^1, I_q^2 ..., I_q^N$. A deep CNN backbone is used in a Siamese network configuration with shared parameters to generate query(ϕ_q) and gallery embeddings(ϕ_g) stored in the gallery. In both cases, ϕ can be either from a single image (image-based ReID) or a batch of image patches of the same object(video-ReID). Multiple

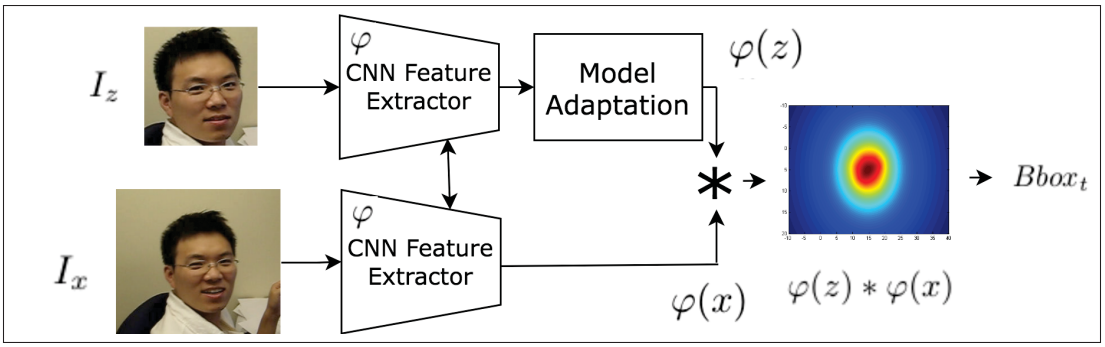

Figure 0.2 Illustration of a general tracker with Siamese architecture. Deep CNN backbones with shared parameters are feature extractors that extract embeddings from object templates and background patches. Object location is determined by correlating the embeddings of the template and background patch

embeddings from the same tracklet (a collection of bounding boxes of an object in a video sequence) are aggregated to generate a single embedding in the case of video-based ReID.

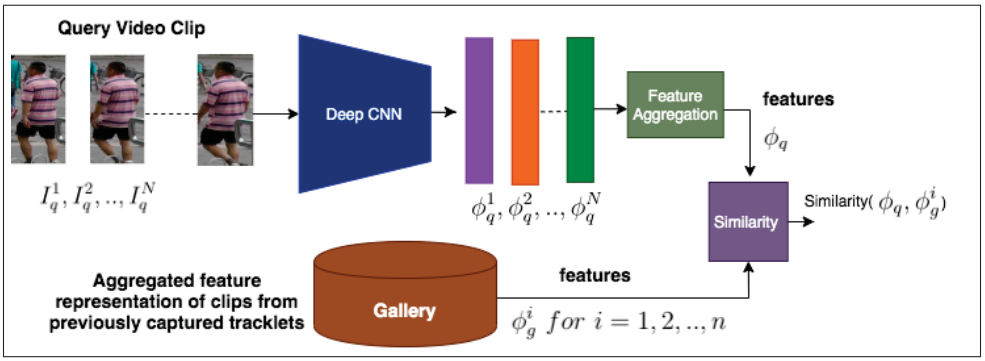

Figure 0.3 Illustration of a general tracker with Siamese networks as the backbone. Deep CNN backbones with shared parameters are feature extractors that extract embeddings from object templates and background patches. Object location is determined by correlating the embeddings of the template and background patch

Image-based ReID is computationally efficient compared to video-based ReID but is more sensitive to challenges such as partial occlusion, localization error, etc.

Challenges: In this section, we discuss some key challenges in tracking and ReID for video surveillance applications.

- **Obtaining Good image sequence (video) Feature Representation**: As discussed in the literature above, learning a good video feature representation from image sequences with Siamese networks is essential for video-based ReID and tracking. Aggregating multiple features using different image samples from a given video into a single feature representation of the object is crucial for matching. In particular, aggregating features from a longer sequence length can be challenging (Wei, Yang, Zuo, Qian & Wang, 2022). (Subramaniam, Nambiar & Mittal, 2019c; Eom, Lee, Lee & Ham, 2021; Zhang, Lan, Zeng & Chen, 2020d) have proposed systems that can handle aggregation of up to 8 video frames into a single sequence and the average number of aggregated frames in literature being 4-6 (Gao & Nevatia, 2018a; Ye *et al.*, 2021). Aggregating greater than 6/8 frames has resulted in lower performance due to the limitation of feature aggregation mechanisms.
- **Partial Occlusion of the Object**: Occlusion is one of the major challenges in tracking (Liu *et al.*, 2022; Jin *et al.*, 2024) and ReID (Zhuo *et al.*, 2018b; Li *et al.*, 2024). With partial occlusion, tracking can begin to drift, and the track will eventually be lost. In ReID, occlusion can cause mismatches. Occlusion is often handled in tracking by the re-detection of the target once the target is completely lost. Especially in adaptive tracking, where the object is learned online, occlusion causes distraction, and the occluder is learned along with the object, causing track drift (Gavves, Tao, Gupta & Smeulders, 2021a). Hence, it is necessary to address this issue within the tracker. Regarding ReID, the occlusion problem has been previously handled using multiscale features (Wang *et al.*, 2020b) and additional cues such as pose (Gao, Wang, Lu & Liu, 2020b), making the overall computational complexity unsuitable for video surveillance applications.
- **Computational Complexity**: Real-time video surveillance applications are constrained by the computational complexity of the overall system. Especially with deep networks used as backbones when handling different sub-tasks like tracking and video surveillance, the overall

complexity is high and limits the usage in real-time applications (Baharani, Mohan & Tabkhi, 2019).

0.2 Research Statement and Contributions

Following the challenges and limitations highlighted above, the main research question addressed in this work is whether a good video representation can be learned for treating tracking and ReID as a matching problem and if the system can be made robust to occlusion with constraints of real-time video surveillance. Specifically, four research directions have been explored for object localization through adaptive tracking and robust object/person re-identification for real-time video surveillance. The core contributions of this thesis are:

Chapter 2: Adaptive Learning from Image Sequences for Object Tracking

In Chapter 2, the challenge of learning a good representation from image sequences has been visited. This has been done in a two-step process. Firstly, efficient online learning of object representation is handled by dynamic template selection and entropy maximization of the samples from an image sequence. Secondly, track drift has been addressed by learning occlusion-aware features. Change detection techniques have been used to detect track drift and occlusions, and sample selection is influenced accordingly, which further helps computationally efficient learning. Contributions of various hyperparameters, as well as their influence on the complexity of the tracker, have been studied.

Related Publications:

Kiran, M., Nguyen-Meidine, L.T., M. O. Cruz, R., Blais-Morin, L.A. and Granger, E., Incremental Template Learning for Occlusion-Aware Visual Object Tracking. Expert Systems With Applications. Submitted.

Kiran, M., Sahay, R., Cruz, R.M.O.E., Blais-Morin, L.A. and Granger, E., 2022, July. Dynamic template selection through change detection for adaptive Siamese tracking. In 2022 International Joint Conference on Neural Networks (IJCNN). IEEE.

Kiran, M., Nguyen-Meidine, L.T., Sahay, R., Cruz, R.M.O.E., Blais-Morin, L.A. and Granger, E., 2022, June. Generative target update for adaptive siamese tracking. In International Conference on Pattern Recognition and Artificial Intelligence Cham: Springer International Publishing.

Kiran, M., Tiwari, V., Morin, L.A.B. and Granger, E., 2019, September. On the interaction between deep detectors and Siamese trackers in video surveillance. In 2019 16th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS). IEEE.

Chapter 3: Representation Learning from Image Sequences for Person ReID : This chapter addresses the problem of representation learning of image sequences for Person ReID applications. Although tracking and ReID are mostly Siamese matching networks, the representation learning can slightly differ because the object changes gradually during tracking, and for ReID, the object can be from a different view. To address this, we model image sequence representation using longer sequences and optical flow for attention, thereby being able to model spatiotemporal features of the person. The outcome of this work was that the system could use significantly longer sequences for person matching, thereby improving the overall performance of ReID.

Related Publication:

Kiran, M., Bhuiyan, A., Blais-Morin, L.A., Ayed, I.B. and Granger, E., 2021. Flow guided mutual attention for person re-identification. Image and Vision Computing.

Chapter 4: Leveraging Dissimilarity Space for Occluded Person ReID: This chapter presents a method for learning computationally efficient representations robust to partial occlusions for ReID applications. In particular, we propose a method to leverage the distributions of holistic (case where the majority of the samples are non-occluded) ReID data to be robust to occlusions

using the dissimilarity space. Our proposed method achieves state-of-the-art performance in Occluded ReID without additional cues such as pose or segmentation information.

Related Publications:

Kiran, M., Praveen, R.G., Belharbi, S., Blais-Morin, L.A. and Granger, E., 2021. British Machine Vision Conference.

Chapter 5: Leveraging Dissimilarity Space for Distance Learning in ReID: This contribution explores the space for dissimilarity in Person ReID to solve the overlap problem between classes observed in Euclidean space. This is worsened by computationally efficient models with smaller representation capacities and limited training data. We propose applying dichotomy transformation to feature space and classifying sample pairs as similar or dissimilar with a max-margin classifier. We then propose to use the distance of samples from the margin as a measure of distance between classes. End-to-end learning of a classifier in the dissimilarity space has been observed to improve retrieval accuracy for Person ReID tasks.

0.3 Organization of the Thesis

This thesis's main text, including articles and chapters 2 to 5, corresponds to journal and conference publications. Chapter 1 presents a brief background on tracking and ReID. This Chapter provides a classification of various tracking and ReID methods and their relevant background. Chapter 2 presents a method for incremental learning in online trackers. Chapter 3 explores the problem of video representation learning in ReID. Person ReID under occlusion is explored in Chapter 4. We discuss the advantage of dissimilarity space and max-margin classifier in Chapter 5. Chapter 6 discusses the contributions and limitations of the thesis and provides additional directions for future work.

CHAPTER 1

TRACKING AND REID IN VIDEO SURVEILLANCE

As discussed in the Introduction, tracking and ReID fit nicely in the narrative of similarity-matching problems. In both applications, there is an exemplar image (target template in tracking and gallery sample in ReID) and a search image (last known target region in the image for tracking and query image for ReID). The exemplar image matches the candidate images to obtain the best match. Similarity matching is an important aspect of pattern recognition. When two vector representations of data are compared, Euclidean distance or suitable distance function is used to calculate their similarity. However, this can quickly get complex to data requiring further processing or compression before comparison with a distance function for similarity matching. In such cases, a Siamese neural network can be a good choice (Chicco, 2021). Siamese neural networks are two identical neural networks capable of encoding and capturing the hidden representation of data. Siamese architecture ensures consistency of predictions (Koch *et al.*, 2015) and are considered promising architectures for similarity matching problems based on their balance between performance and efficiency (Ondrašovič & Tarábek, 2021). This Chapter focuses on mainstream tracking and ReID methods using Siamese networks. We review neural networks and deep learning, and then we review Siamese networks and architectures. In the following section, we discuss state-of-the-art tracking and ReID models. Finally, we discuss the challenges, limitations, and potential opportunities with the Siamese network for video surveillance applications.

1.1 Deep Learning

Deep learning (DL) is a subset of machine learning (ML) inspired by information processing patterns in the human brain. DL does not require handcrafted features to be designed by a human operator but uses large amounts of data to map a given input to label. With deeper feature representations, DL models can capture complex patterns in data where the representation is optimized for a given task or application. With the transfer learning approach, a model trained for an application can be fine-tuned in another application or task. Also, DL approaches are

highly scalable. Scalability can be explained in different contexts as 1) Data Scalable, Ability to learn from large amounts of data; 2) Compute Scalable: Large DL models can be scalable by taking advantage of GPU and TPUs given the application constraints and bounds; 3) Scalable to Diverse domains: Some applications, such as real-time object detection and autonomous driving, require low-latency and real-time processing. DL models can be optimized for such applications.

1.1.1 Taxomomy of Deep Networks

DL approaches can be classified as supervised, semi-supervised, and unsupervised learning. Deep Supervised Learning techniques deal with labeled data. A DL network uses a collection of data (X, y) where X is the input data and y is the label. The network uses the data to learn to predict $\hat{y}_t = f(X_t)$ where $_t$ represents test time data. Various architectures, such as Recurrent Neural Networks(RNN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Deep Neural Network (DNN), are used for learning. Semi-supervised learning is an approach that relies on utilizing semi-labeled datasets as a foundation for the learning process. At times, generative adversarial networks (Dai, Yang, Yang, Cohen & Salakhutdinov, 2017) (GANs) and deep reinforcement learning (Dong, Xia & Peng, 2021) (DRL) are employed similarly to this technique. Furthermore, recurrent neural networks (RNNs), including Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, are also utilized for partially supervised learning. One notable advantage of this method is its ability to reduce the requirement for extensively labeled data. One drawback of this approach is the potential for irrelevant input features in the training data to introduce inaccuracies. Deep Unsupervised learning makes it possible to learn from data without labels. The system learns interior representation and data patterns to discover the unidentified relationships in the data. This technique includes dimensionality reduction and clustering. Several DL techniques perform well in dimensionality reduction and clustering (Caron, Bojanowski, Joulin & Douze, 2018), including Boltzmann machines (Fischer & Igel, 2012), Autoencoders (Krizhevsky & Hinton, 2011), and GANs (Creswell *et al.*, 2018).

1.1.2 Neural Network and Deep CNN Architectures

The McCulloch-Pitts neuron is a simplified mathematical model of a biological neuron proposed by Warren McCulloch and Walter Pitts (McCulloch & Pitts, 1943). McCulloch-Pitts neuron was modeled as a binary threshold unit. The model performs repeated computations in time steps$t = 0, 1, 2, 3, ...$ The state of neuron j at time step t is given by $s_j(t)$,

$$s_j(t) = \begin{cases} -1 \text{ inactive} \\ \\ 1 \text{ active} \end{cases} \tag{1.1}$$

Given the states $s_j(t)$, the neuron i computes,

$$s_i(t+1) = \text{sgn}\left(\sum_{i=1}^{N} w_{ij} s_j(t) - \theta_i\right) \equiv \text{sgn}\left[b_i(t)\right] \tag{1.2}$$

sgn is the signum or thresholding function.

$$\text{sgn}(b) = \begin{cases} -1, & b < 0, \\ \\ +1, & b \geq 0. \end{cases} \tag{1.3}$$

The argument of the sgn function b is defined as,

$$b_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t) - \theta_i. \tag{1.4}$$

In Eqn. 1.4, a small change in when b is close to zero, a small change in the inputs of Eqn. 1.2 can cause large fluctuations in the activity levels of the neuron. Such fluctuations can be dampened by allowing the neuron to respond continuously rather than discretely. Eqn. 1.2 can be modified

as below,

$$s_i(t+1) = g\left(\sum_j w_{ij} s_j(t) - \theta_i\right) \tag{1.5}$$

In Eqn. 1.5, $g(b)$ is a continuous activation function. $g(b)$ can be either linear or non-linear.

Convolutional networks Although convolutional neural networks(CNN) have been around since the 1980s (LeCun *et al.*, 1989), they attracted much attention with the success of ImageNet Challenge (Krizhevsky, Sutskever & Hinton, 2012b). The main difference with multi-layer perceptrons is that CNNs accept inputs with a spatial array of input terminals. Secondly, these layers act like local feature detectors like edge and corner detection, especially for image-based applications. They map an input to an output called a feature map. CNNs can incorporate multiple convolutional layers. In addition to feature maps, CNNs encompass various layers, such as normalization like BatchNorm (Ioffe & Szegedy, 2015). For instance, pooling layers execute local averaging operations on the outputs from convolution layers, aiming to expedite the learning process by reducing the number of parameters. Furthermore, convolutional networks may also include fully connected layers.

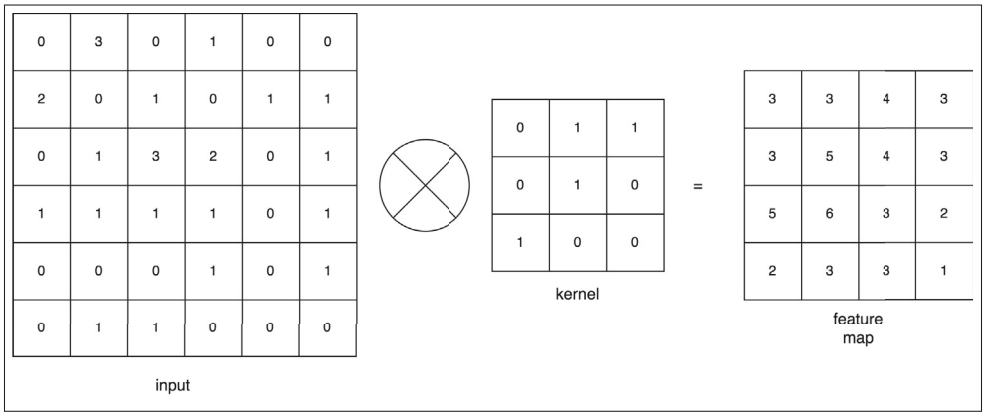

Figure 1.1 Illustration of convolution operation on a sample input. A kernel of size 3×3 is used with stride 1

Fig. 1.1 shows an example of convolution operation on a 2D input array data. The kernel forms the spatial array of input terminals with weights. A kernel is placed at the top left corner to calculate the response and then performs a row-wise and column-wise element multiplication

followed by summation. The next calculation is performed by moving the filter or kernel by stride to the right. In Fig. 1.1, a stride of 1 was used. This is continued with downward vertical movement of the filter when the leftmost side of the input array is reached, and this is continued until reaching the bottom of the input.

Convolutional Neural Networks (CNN) are commonly used architectures in Deep Learning (Huang, Liu, Van Der Maaten & Weinberger, 2017; Krizhevsky *et al.*, 2012b). The main benefit of CNN compared to its predecessors is its ability to identify features for various tasks automatically (Gu *et al.*, 2018). (Goodfellow, Bengio & Courville, 2016) have outlined three principal advantages of Convolutional Neural Networks (CNN): equivalent representations, sparse interactions, and parameter sharing. In contrast to traditional fully connected (FC) networks, CNNs leverage shared weights and local connections to fully exploit the inherent 2D structures of input data, such as image signals. This approach utilizes a remarkably compact parameter set, simplifying the training process and accelerating network performance. This phenomenon aligns with the functioning of cells in the visual cortex, where only small regions of a scene are perceptually processed, as opposed to the entire scene. In essence, these cells spatially extract local correlations present in the input, akin to the operation of local filters over the input.

The history of CNN starts with LeNet, mainly used for handwritten digit recognition applications (LeCun *et al.*, 1995). AlexNet (Krizhevsky *et al.*, 2012b) was a breakthrough in deep CNN by winning the ImageNet ILSVRC Challenge (Russakovsky *et al.*, 2015b). AlexNet introduced the ReLU activation function to solve for vanishing gradients. AlexNet made a significant impact on the later CNN generations. With the success of AlexNet, efficient and systematic design principles were proposed by (Simonyan & Zisserman, 2014b) to develop VGGNet. VGGNet is a multi-layer model with a larger number of layers than AlexNet. VGGNet showed experimentally that parallel assignment of smaller-sized filters in the convolutional layers is more efficient and effective than large-size convolutional filters. VGGNet achieved significant performance improvement in localization and classification tasks compared to its predecessors. GoogleNet (Szegedy *et al.*, 2015), also called Inception V1, emerged as the winner of the ILSVRCC 2014 challenge. GoogleNet used multi-scale convolutional functions to improve the

efficiency of CNN parameters and enhance learnability. Efficient 1x1 bottle-neck layers were introduced ahead of filters with large kernel sizes. ResNet (He, Zhang, Ren & Sun, 2016) was the winner of the 2015 ILSVRC challenge. The main challenge addressed in ResNet's design was to design a model free of vanishing gradients with an ultra-deep CNN design. This was achieved by skipping connections between layers. Hence, this is a conventional feed-forward network integrated with residual connections. DenseNet (Huang *et al.*, 2017) improved upon the design of ResNet with cross-layer connectivity. The network uses a feed-forward approach where each layer is connected to all the layers in the network. DenseNet concatenates the features from different layers rather than adding them.

2D CNNs were extended to 3D CNNs (Tran, Bourdev, Fergus, Torresani & Paluri, 2015) to apply the CNN to model spatio-temporal information. 3D CNNs were applied to video data for various tasks such as action recognition (Ji, Xu, Yang & Yu, 2012), video classifications (Diba, Pazandeh & Van Gool, 2016), etc. However, 3D CNNs fail to capture long-term temporal patterns due to the complexity of 3D CNNs.

1.1.3 Siamese Networks

1.1.3.1 Introduction and early works

Siamese networks were introduced by (Bromley, Guyon, LeCun, Säckinger & Shah, 1993) for signature verification application. The Siamese neural network is a twin network that accepts two inputs but is joined at the highest level feature output layer by a distance function or energy function. The twin networks share the neural network parameters. This ensures that two inputs that are very similar to each other will be mapped to the same neighborhood. Also, weight sharing makes the network symmetric (Koch *et al.*, 2015).

However, in the work by (Chopra, Hadsell & LeCun, 2005), a contrastive energy function was employed, featuring dual terms aimed at reducing the energy associated with similar pairs and augmenting the energy related to dissimilar pairs. (Koch *et al.*, 2015) adopt a methodology

incorporating the weighted L1 distance between the paired feature vectors, $h1$ and $h2$, extracted from the neural network while employing a sigmoid activation function. This sigmoid activation maps the resulting values to the bounded interval $[0, 1]$. As opposed to an energy function, (Koch *et al.*, 2015) showed that using a cross-entropy function for learning is more suitable for a siamese setting. Fig. 1.2 illustrates a simple siamese network citep (Koch *et al.*, 2015) with logistic prediction p. The prediction vector is given as $p = \sigma\left(\sum_j \alpha_j \left|\mathbf{h}^{(j)}_{1,L-1} - \mathbf{h}^{(j)}_{2,L-1}\right|\right)$. This is for a generic neural network where the features from the $(L-1)$th layer are used to calculate the distance. α_j are additional parameters learned during training weighing component-wise distance between the feature vectors. Hence, with their improvements to the method of (Chopra *et al.*, 2005), (Koch *et al.*, 2015) show that with their proposed training, they could leverage the discriminative power of siamese networks not just to new data but to entirely new classes of unknown distributions.

1.1.3.2 Embedding and Classification Losses

Pair and triplet loss functions form the basis for two pivotal strategies in metric learning (Musgrave, Belongie & Lim, 2020). An established pair-based technique is exemplified by the contrastive loss (Chopra *et al.*, 2005), which strives to minimize the distance between positive pairs (d_{pos}) below a predefined threshold (m_{pos}) and maximize the distance between negative pairs (d_{neg}) beyond a specified threshold (m_{neg}):

$$L_{\text{contrastive}} = \left[d_{\text{pos}} - m_{\text{pos}}\right]_+ + \left[m_{\text{neg}} - d_{\text{neg}}\right]_+ \tag{1.6}$$

The triplet loss function introduced by (Weinberger & Saul, 2009) comprises three elements (embeddings): an anchor ϕ_a, a positive sample ϕ_p, and a negative sample ϕ_n, with the anchor being inherently more akin to the positive instance than the negative one. The triplet margin loss shown below is designed to minimize the distances between the anchor and positive samples (d_{ap}), striving to ensure that they are smaller than the distances between the anchor and negative

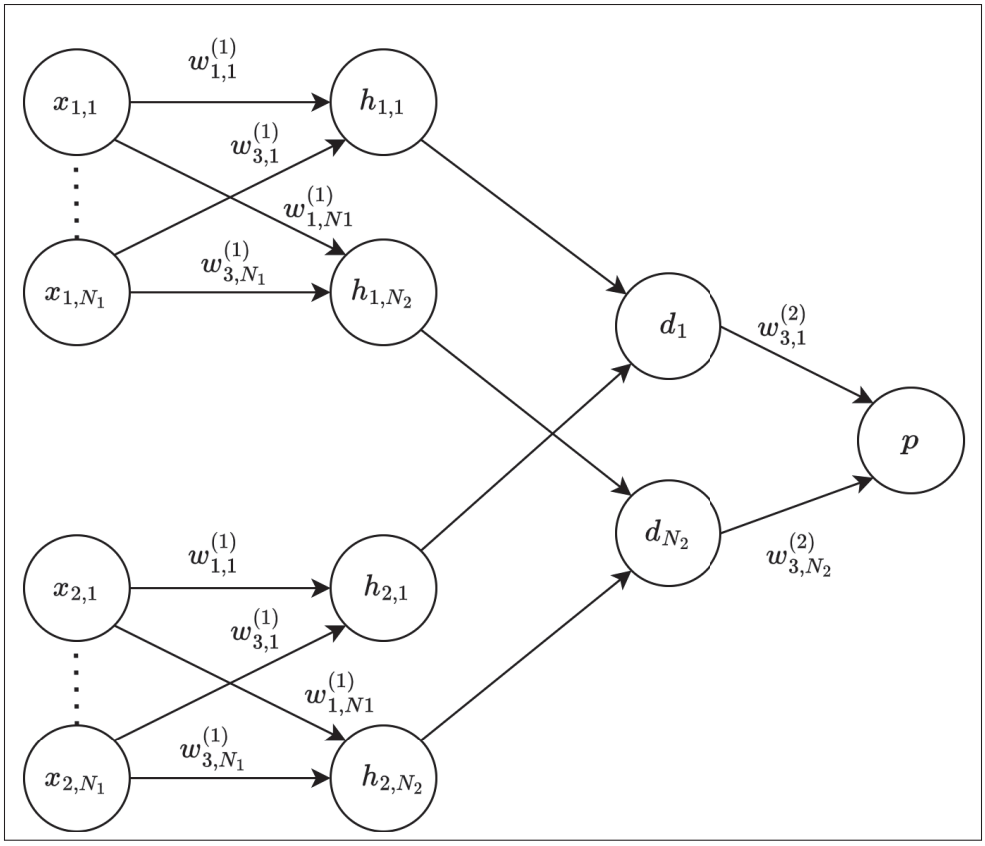

Figure 1.2 A simple illustration of siamese network with logistic prediction indicated by p. The network is a two-layer twin network with shared weights, as shown. Note that vector d is formed by applying a distance function between the hidden representations of the two inputs. The matrix w represents the shared weights between the twin networks
Taken from Koch *et al.* (2015)

samples (d_{an}) by a pre-established margin (m)+

$$L_{\text{triplet}} = \left[d_{a_p} - d_{a_n} + m\right]_+ \tag{1.7}$$

Classification losses are constructed by incorporating a weight matrix, each column corresponding to a specific class. Typically, the training process involves multiplication between the weight

matrix and embedding vectors to derive logits, followed by a designated loss function.

$$\mathcal{L}_{\mathcal{S}'} = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{s\tilde{W}_{y_i}^{T}\tilde{\phi}_i}}{\sum_{j=1}^{n}e^{s\tilde{W}_{j}^{T}\tilde{\phi}_i}} \tag{1.8}$$

Normalized softmax loss is a straightforward example for this introduced by Wang, Xiang, Cheng & Yuille (2017), which is identical to cross-entropy but with the weight matrix of the classifier, $L2$ normalized. This is shown in Eqn. 1.8 where $\tilde{\phi}_i$ represents a normalized feature of a i^{th} sample and $\tilde{W}$ the normalized weight matrix and s represents the scale. As opposed to this, ProxyNCA applies Euclidean distances instead of cosine similarities to the weight matrix (Movshovitz-Attias, Toshev, Leung, Ioffe & Singh, 2017).

In addition to using the loss functions described above, additional information is often learned by architectural changes to Siamese networks. (Zhang *et al.*, 2023b) learn global information with Siamese networks for remote sensing applications. Attention with Siamese networks has been explored by Yin *et al.* (2023) for change detection applications. (Valero-Mas, Gallego & Rico-Juan, 2024) have used Siamese networks in ensemble learning approaches for few-shot classification.

1.1.3.3 Pair Mining

Several pairs of samples can be generated for the training siamese network since the samples are classified as similar or dissimilar. All feasible pairs or triplets present notable drawbacks to training a Siamese network. In practical terms, when trained on easier examples, it can lead to substantial memory consumption and cause overfitting.

Mining is finding the best pairs to train to make the model generalizable. Mining can be done either online or offline. Offline mining involves forming pairs before the start of training and having access to this information during training. The online mining process aims to identify challenging pairs or triplets within each randomly sampled batch. Theoretically, it tends to encompass a substantial number of relatively straightforward negative and positive instances, which, in turn, can result in a rapid plateauing of performance. Consequently, (Hermans, Beyer & Leibe, 2017a)

propose an intuitive approach to focus exclusively on the most challenging positive and negative samples. However, it has been observed that this method can lead to the emergence of noisy gradients and convergence towards unfavorable local optima (Yu, Liu, Gong, Ding & Tao, 2018).

Conversely, (Wu, Manmatha, Smola & Krahenbuhl, 2017a) have demonstrated that semihard mining tends to yield limited improvements as the count of semihard negatives decreases. They assert that distance-weighted sampling leads to diverse negative examples, encompassing easy, semihard, and hard instances, enhancing performance. Furthermore, online mining can be seamlessly integrated into the architectural framework of models. In a recent study, (Wang, Han, Huang, Dong & Scott, 2019) introduced an uncomplicated pair mining strategy. Negative samples are selected if they are closer to an anchor than its most challenging positive instance. In contrast, positive samples are chosen if they are farther from an anchor than its most challenging negative instance. (Xuan, Stylianou, Liu & Pless, 2020) have shown that hard negative samples, although useful, suffer from optimization issues. They further propose gradient normalization methods to fix the optimization issues. It can be observed that sample mining is still an open problem and needs careful addressing to help select optimal pairs for training siamese networks. Some other works in this domain include negative sample mining (Wang, Wang, Wu, Li & Wu, 2022b), where the authors emphasize the importance of negative samples for metric learning. As proposed to traditional mining strategies, (Yan, Luo, Deng & Huang, 2023) have proposed hierarchical mining, which is insensitive to noisy labels.

1.2 Visual Object Tracking

The Visual Object tracking algorithms can be classified as shown in Figure 1.3. Visual Object Tracking methods can be classified either based on their type of classification or based on detection with data association methods, which is often the case in Multi-Object tracking.

Many trackers have a fixed appearance model throughout their life cycle while tracking an object. Example Siamese Fully Convolutional tracker (Bertinetto, Valmadre, Henriques, Vedaldi & Torr, 2016) uses a single object representation. At the same time, many other methods choose to

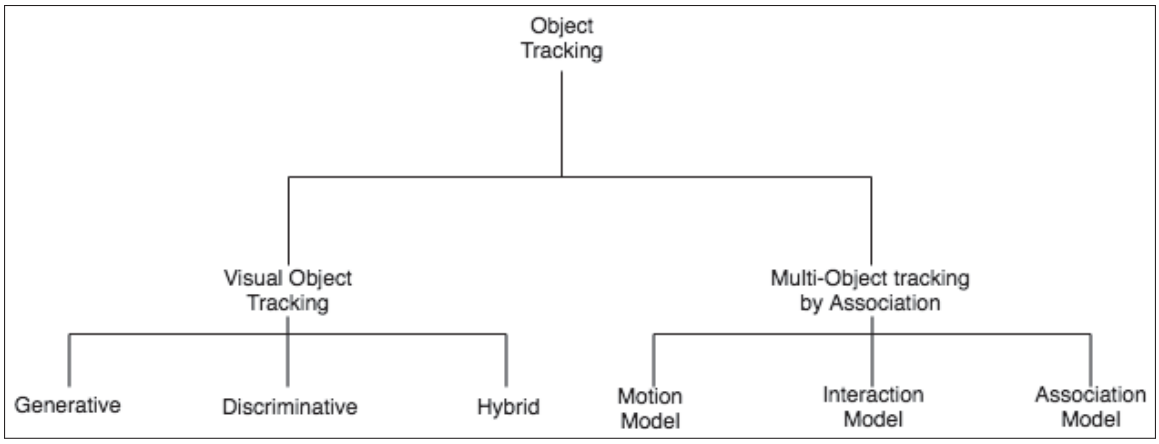

Figure 1.3 Classification Of Visual Object tracking algorithms

update the model of their representation using various strategies, which remains a key to an adaptive tracker.

Visual Object trackers are local object detectors that learn the object online and locally search for the object in a video stream. They often have a classifier module, and the tracker can be classified based on the type of classification. They often have Generative Trackers localize the object by only considering the maximum likelihood that relies only on the object's appearance. Generative trackers often use sparse features and are much simpler in construction. Discriminative trackers discriminate the object's appearance from that of the background with a classifier that has learned a decision boundary between them. Hybrid trackers use a combination of both to work hand in hand. For example, a Generative algorithm could be initially used to produce samples for the Discriminative part to start learning the discriminative boundary.

STRUCK (Hare *et al.*, 2016a) is a discriminative tracker approach proposed to avoid the accumulation of tracking errors. The feature extraction method chooses grey-level image features, Histogram of Orientation features, etc. The classifier proposed is a Structured Output classifier. The system uses a Structured output SVM to learn the object transformation between frames, $f : X \rightarrow Y$, instead of predicting $\pm$, the label. f is a learned in a Structures Output SVM framework which introduces a $F : X \times Y \rightarrow R$ to predict

$$y_t = f(x_t^{p_{t-1}}) = argmaxF_{y\epsilon Y}(x_t^{p_{t-1}}, y) \tag{1.9}$$

To update the prediction function online, a labeled example is supplied relative to the new tracker location $(x_t^{p_t}, y^0)$.

One major advantage of the STRUCK tracker is that it selects the samples based on their current state and labels them for training in a Structured Output SVM model.

Kernelised Correlation Filter is a discriminative tracker (KCF) (Wang, O'Brien, Xiang, Xu & Najjaran, b). The general correlation filter-based tracking framework can be summarised as follows. Initially, an image patch with the object is selected. The filter is trained based on this patch in the first frame. Various features are extracted, and correlation operation is performed by piece-wise multiplication using Discrete Fourier Transform (DFT) instead of convolution in the space domain. As a result of the correlation process, a spatial confidence map is obtained after an inverse Fast Fourier Transform (FFT). The new state of the target is predicted from the position of the maximum value in the map. The estimated position is then used to update the model and correlation filter. The authors suggest various feature extraction methods, including greyscale image intensities, HOG, etc., could be used.

$$x * h = F^{-1}(\hat{x'} \cdot \hat{h*}) \tag{1.10}$$

KCF uses a circulant matrix instead of translated patches around the target region(commonly used in a conventional tracker). The system solves a Ridge Regression problem given training patterns and labels $(x_1, y_1), (x_2, y_2), .., (x_m, y_m)$, a classifier $f(x)$ is trained that minimize the regularized risk.

L is the loss function and λ is the regularization term. When the kernel trick is applied to the above minimization problem, it will also help to work with high-dimensional features. The

solution of the above equation in closed form is

$$\alpha = (K + \lambda I)^{-1} y \tag{1.11}$$

If X is a $1 \times n$ patch denoting a feature vector (base sample), the goal is to use the base sample and generate several virtual samples by translating it. The one-dimensional translation of this vector can be modeled by a cyclic shift operator using P, as shown below. Hence, each row of P translates x by one element.

$$P = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{1.12}$$

The innovation of the KCF tracker is that the shifted samples are used as input to solve the ridge regression problem. First, a set of shifted samples is created, as shown below.

$$X = C(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_n & x_1 & x_2 & \cdots & x_{n-1} \\ x_{n-1} & x_n & x_1 & \cdots & x_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_2 & x_3 & x_4 & \cdots & x_1 \end{bmatrix} \tag{1.13}$$

X is a circulant matrix with one interesting property, i.e., a circulant matrix can be diagonalized with the Fourier Transform. This can be expressed as,

$$X = F \operatorname{diag}(\hat{\mathbf{x}}) F^H \tag{1.14}$$

GOTURN (Held, Thrun & Savarese) tracker is a generative tracker that uses no discrimination against the background but regresses the object's location directly via a Deep Siamese CNN.

CaffeNet architecture is used as the feature extraction layer. The fully connected layer in the original CaffeNet is excluded, and the results from the POOl5 layer of the CaffeNet architecture are used to represent the target as illustrated in figure 1.5

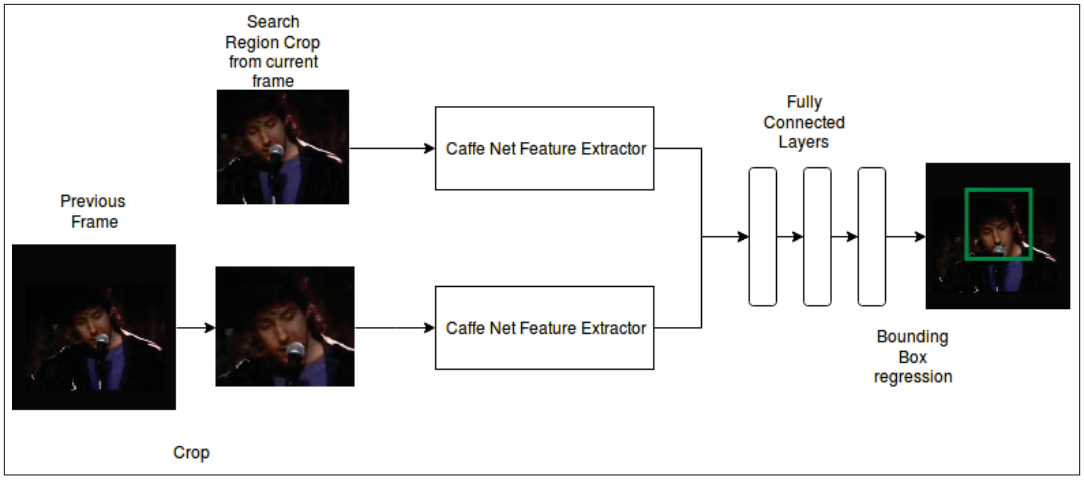

Figure 1.4 Architecture of GOTURN Object tracking algorithm

The GOTURN architecture has three fully connected layers that accept input from the feature extraction layers. The extracted features from the target and search regions are fed to the Fully Connected layer. The fully connected layer has been trained to compare the features from the two images and regress the location coordinates of the target image in the search region.

In the first frame of the video, a target region is specified, and the target region of interest is further padded with background information in the full image. The search region in the next frame will be the same as the coordinates of the region of interest (ROI) of the target image from the previous frame. The Fully connected layers regress the location coordinates of the target in the current frame. This ROI is centered in the image and padded to become the target image for the next frame. Hence, the Target image is recursively updated based on the results of the previous frame. One important innovation in the GOTURN tracker is that regardless of using deep CNNs that normally require training with the object to be tracked, they use only generic videos for training to learn the transformation of the template between two consecutive frames. The network is trained on an ALOV300++ video tracking dataset. Training involves selecting successive pairs of images from video. The images are then cropped per the target size

in $(n-1)th$ frame and padded with the target in the center. The search region is defined as the ROI of the target in $(n-1)th$ extracted in the nth frame. Ground truth is modified to represent the target coordinates in the nth frame based on the frame of reference of the search region.

ROLO Tracker (Ning *et al.*, 2017) is an example of a Hybrid tracker that uses a discriminative model to detect objects in real time while using Spatiotemporal information to track the model across frames using an LSTM layer.

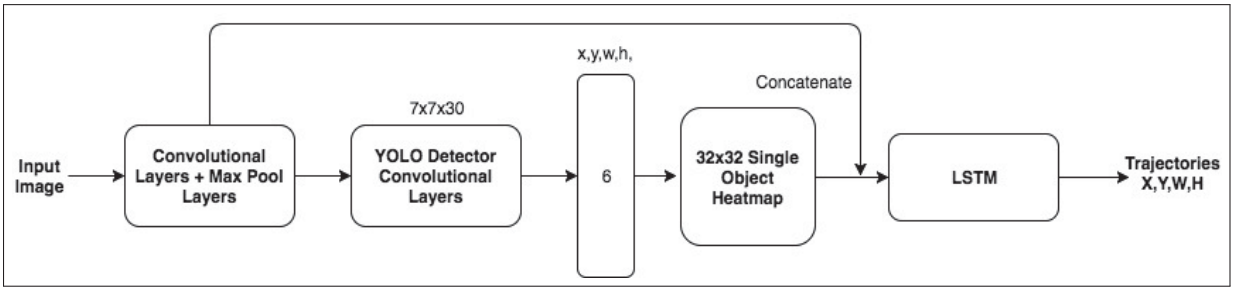

Figure 1.5 Architecture of the ROLO tracker

GoogLENet architecture is used as the feature extractor. The fully connected layer is removed from the original GoogLENet architecture, and the polling layer before the fully connected layer serves as the output node for the feature extractor. The extracted features are high dimensional convolutional features. The YOLO detector uses these features to detect objects and extract bounding boxes. The LSTM-based tracker also uses these features to track the object detected by YOLO. The YOLO (Redmon & Farhadi, 2016) object detector handles the target detection and initialization of the tracker. YOLO object detector is a single-pass object detector that uses a regression-based technique to detect objects in images. The original YOLO was trained on a twenty-class VOC dataset. The output of the YOLO detector is a tensor of size six that includes center coordinates, width, height, class confidence, and detector confidence. The detector and class confidences are skipped, and the bounding box coordinates are used as input to initiate and update the tracker. The detection is performed in every frame, and the tracker tracks and associates the object between frames to form trajectories. A 32x32 heat map represents the object's coordinates to initiate the tracker. The LSTM layer accepts input from the Heatmap. It has been trained to learn spatio-temporal relationships between image features and object

motion. ROLO is a single object tracker; hence, during operation, YOLO detects objects such as a human or a car in the image and tracks the location of the same object in the next frame. The tracker and detector have shared the convolutional layers, forming an example of a unified tracking system. LSTM associates the object detected in the last frame to the current frame. The detector is called every frame. The system has been trained to have a mean squared error loss on the predicted Heatmap and Ground truth Heatmap.

$$L_{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left\| H_{\text{target,i}} - H_{\text{pred,i}} \right\|_2^2 \tag{1.15}$$

Where n is the number of samples in the batch, this tracker's interesting point is that heatmap-based tracking could be easily extended to a Multi-Object case as all the objects can be represented on a single heatmap.

Siamese Fully Convolutional trackers (Bertinetto *et al.*, 2016) (SiamFC) are discriminative trackers. The architecture is somewhat similar to the GOTURN architecture. They have two CNN feature extraction layers. The inputs to these layers are a template image and a search region in the image stream with an area two times that of the template image. In a frame n, the search region has been extracted from the tracker output at time frame $n - 1$ around a region padded around the location of the object found by the tracker at frame $n - 1$. But unlike GOTURN, where the template is updated every frame, in SiamFC, the template remains the same from the very first frame of the object.

During runtime, the features of the template region are convolved over those of the search region to produce a heatmap. The location of maxima in the heatmap indicates the object's location within the search region. Before the 'argmax'(get the location of maximum value) operation, the heatmap is scaled to the input image size with Bi-Cubic interpolation to extract the location in image space.

The system is trained in a discriminative fashion. Since the search region is the template with some padded background information, the system is trained to discriminate the template with the background. The large search images and training with a logistic loss facilitate discriminative

learning

$$\ell(y, v) = \log(1 + \exp(-yv)) \tag{1.16}$$

v is the real-values score of a single exemplar-candidate pair and $y \in \{+1, -1\}$. During training, the fully convolutional nature is exploited to generate exemplar, and large search region pair in both cases object-centered to produce a map of scores $v : \mathcal{D} \rightarrow \mathbb{R}$

$$L(y, v) = \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} \ell(y[u], v[u]) \tag{1.17}$$

Hence, the Ground truth is a virtual score map generated by applying the following rule.

$$y[u] = \begin{cases} +1 & \text{if } k\|u - c\| \leq R \\ -1 & \text{otherwise} \end{cases} \tag{1.18}$$

The score map is considered positive if it is within a radius R of the center, and k is the network's stride. Since the system is fully convolutional, it can track at a speed of 100 frames per second during tracking. It is important to note that further research in tracking followed the SiamFC paradigm and overall architecture with improvements, including architecture changes, online learning, and treating tracking as regression and classification problems, as described below.

SiamRPN (Li, Yan, Wu, Zhu & Hu, a) further improves this work by employing region proposals to produce a target-specific anchor-based detector. Then, the following Siamese trackers mainly involved designing more powerful backbones (Zhang & Peng, 2019; Li & Zhang, 2019) or proposal networks, like in (Fan & Ling, 2019).

Most recent works on deep CNN-based trackers employ the learning of an online classifier efficiently suitable for real-time applications. (Danelljan, Bhat, Khan & Felsberg, 2019) employs an IOU-Net (Jiang, Luo, Mao, Xiao & Jiang, 2018) like network, which scores an input sample to estimate the overlap, thereby adapting the overlap maximization approach. DiMP (Bhat *et al.*, 2019b) further extends this by employing a model prediction network and a bounding box maximization strategy. They learn a model is online for tracking with minimum optimization

steps. PrDiMP (Danelljan, Gool & Timofte, 2020) further extends the work of (Bhat *et al.*, 2019b), where they track the target by estimating the conditional probability density of the target state given an input image. Unlike DiMP (Bhat *et al.*, 2019b), which uses a confidence value for probability estimation, the conditional probability density of the target state serves as a more precise and direct estimate of the target state.

Some of the latest literature in VOT has explored improved architectures and feature extraction for matching. SimTrack (Chen *et al.*, 2022a) explored transformers for improved feature extraction. In particular, they improve feature interaction of target representational and search region representation by treating them as concatenated features instead of two separate branches commonly used in Siamese architectures (Bertinetto *et al.*, 2016). SeqTrack (Chen, Peng, Wang, Lu & Hu, 2023) further improves upon SimTrack by using a transformer with an autoregressive approach for bounding box prediction, avoiding the design of complicated regression heads.

1.3 Person Re-Identification

Person ReID is a person retrieval problem across non-overlapping cameras. Given a query image of a person, the task is to identify if this person has appeared in another place at a different time captured by another camera. The main challenges in ReID are different viewpoints of the camera (Karanam, Li & Radke, 2015), background clutter (Song, Huang, Ouyang & Wang, 2018a), and occlusions (Huang, Li, Zhang, Chen & Huang, 2018b), (Hou *et al.*, 2019a).

Fig. 1.6 shows the overall person retrieval or person ReID system based on Deep CNN models. It consisted of a gallery of deep CNN features extracted from cutouts of previously seen persons in different cameras. Given a query image cutout of a person, a deep CNN feature extractor extracts features from the cutout, which is matched against the ones in the gallery by calculating a similarity or distance function. The best match is selected by calculating the closest distance with the query. Since the deep CNN feature extractor is shared between the query and gallery images, the network has been used as a Siamese network.

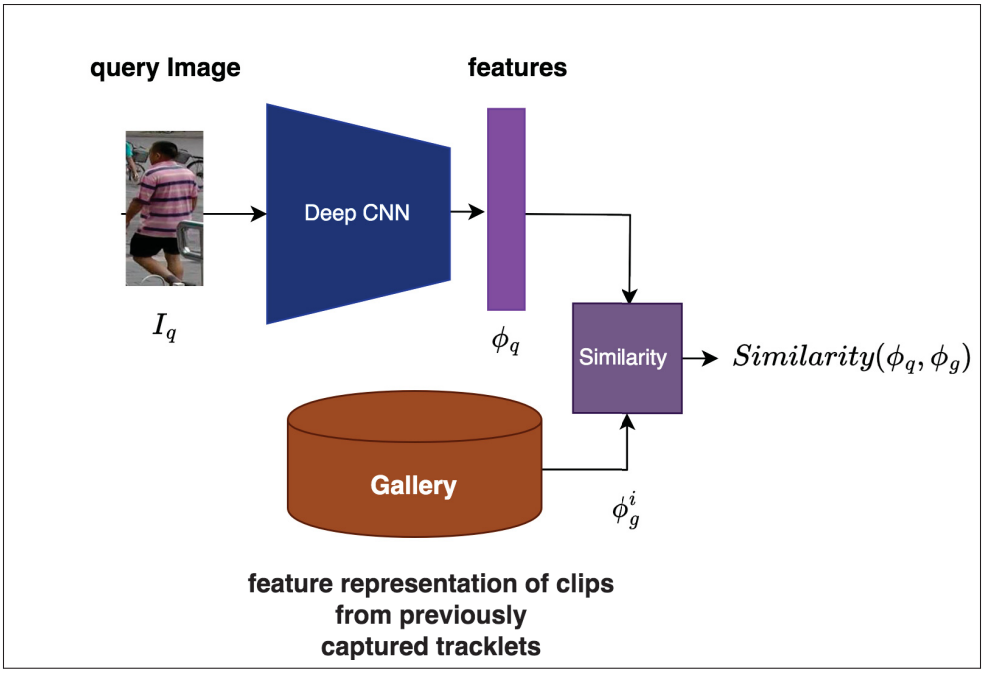

Figure 1.6 Block diagram of a generic DL model illustrating Person-ReID. A query image cutout of a person pre-processed with a person detector is an input to a deep feature extractor. The extracted features are matched against a gallery of previously extracted reference features. The identity of the query is determined based on the ID of the best-matching gallery feature

1.3.1 Image Based ReID

In 2014, both (Yi, Lei, Liao & Li, 2014) and (Li, Zhao, Xiao & Wang, 2014b) introduced the concept of employing Siamese neural networks for person re-identification. Their approach involved training the network to ascertain whether a pair of input images pertain to the same individual. Subsequently, over the ensuing years, numerous methods for person re-identification that utilize Convolutional Neural Networks (CNNs) have been proposed Zhao, Li, Zhuang & Wang (2017c). This section categorizes these CNN-based person-identification methods into distinct groups based on the label information used in training the re-identification models. These categories include fully supervised person re-identification, unsupervised domain-adaptive person re-identification, semi-supervised person re-identification, weakly supervised person re-identification, unsupervised person re-identification, and unsupervised tracklist learning person re-identification. However, it is worth noting that there is a notable scarcity of work

in the remaining groups except for fully supervised person re-identification and unsupervised domain-adaptive person re-identification. Consequently, these less-represented methods are collectively categorized as "others."

Person-ReID for still images has made significant strides in recent years, especially after the introduction of DL networks. Using CNNs for ReID stems from Siamese Network (Bromley, Guyon, LeCun, Säckinger & Shah, 1994), which involves two sub-networks with shared weights and is suitable for finding the pair-wise similarity between query and reference images. It was first used in (Yi *et al.*, 2014) and employed three Siamese sub-networks for deep feature learning. Since then, many authors have focused on designing various DL architectures to learn discriminative feature embedding. Most of these deep-architecture-based ReID (Ahmed, Jones & Marks, 2015; Cheng, Gong, Zhou, Wang & Zheng, 2016; Chen, Chen, Zhang & Huang, 2017a; Liu, Feng, Qi, Jiang & Yan, 2017a; Varior, Haloi & Wang, 2016b) approaches introduce an end-to-end ReID framework, where both feature embedding and metric learning have been investigated as a joint learning problem. In (Ahmed *et al.*, 2015; Varior *et al.*, 2016b), a new layer is proposed to capture the local relationship between two images, which helps model pose and viewpoint variations in cross-view pedestrian images. A few attention-based approaches for deep ReID (Li, Chen, Zhang & Huang, 2017a; Su *et al.*, 2017; Zhao, Li, Zhuang & Wang, 2017b) address misalignment challenges by incorporating a regional attention sub-network into a base re-ID model.

Unlike the previously described literature, attempts were made to extract body part-level features for matching in ReID. Relation network (Park & Ham, 2020), in addition to extracting part-level features, also captures the relation between parts for robust representation learning. Occlusion was identified as one of the important challenges in ReID (Gao, Wang, Lu & Liu, 2020a), and pose estimation was used to improve attention to visible regions. This partially solved the occlusion problem, while the solution was computationally expensive. Further research in ReID focussed mostly on improved feature extraction, like using transformers (Chen *et al.*, 2022b) and fine-grained features with global representation representation (Somers, De Vleeschouwer & Alahi, 2023b).

1.3.1.1 Video Based ReID

Video ReID has recently attracted interest since temporal information allows dealing with ambiguities such as occlusion and background noise (Gao & Nevatia, 2018b; Gu, Ma, Chang, Shan & Chen, 2019; Subramaniam, Nambiar & Mittal, 2019b; Hou *et al.*, 2019b; McLaughlin, d. Rincon & Miller, 2016a). An important problem in video-based ReID is aggregating the image-level features to obtain one single composite feature or descriptor for a video sequence. (Gao & Nevatia, 2018b) have approached this problem by frame-level feature extraction and temporal fusion by using recurrent NNs (RNNs), average pooling, and temporal attention (based on image features). Average Pooling in the temporal dimension can be viewed as summing the features of the sequence by giving equal and normalized weights to them. Average pooling of image instance features from a given sequence has proved useful in most cases, even compared to other DL models based on RNNs or 3D-CNN (Gao & Nevatia, 2018b). 3DCNN has been experimented with in (Gao & Nevatia, 2018b; Li, Zhang & Huang, 2019c) but has not effectively summarised video sequences for ReID. However, there could be some instances of individual images in a sequence that either have higher noise content, or the image's appearance does not contribute much to an individual's identity. These become the debatable cases for Average Pooling. The work of (Eom *et al.*, 2021) identifies distractors in spatial and temporal domains and proposes spatiotemporal memory networks for video person ReID feature extraction and aggregation. Parallel to this work, (Zhang *et al.*, 2021) introduces a spatial-temporal transformer to take advantage of transformer architecture for video feature representation. The importance of sequence length has been emphasized in the work of (Davila *et al.*, 2023), where they introduced a new dataset with extended videos instead of the datasets used in previous methods.

1.3.1.2 Loss functions for Person ReID

The three widely studied loss functions in ReID are identity loss function, verification loss, and triplet loss.

Identity loss treats the ReID problem as a classification problem. A fully connected layer or an MLP is often used after the feature extractor or CNN backbone with output nodes equal to the number of classes or identities. Given a n input image x_i and label y_i, the predicted probability being $p\left(y_i \mid x_i\right)$ then the identity loss is given by,

$$\mathcal{L}_{id} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i \mid x_i\right)\right) \tag{1.19}$$

In Eqn 1.19 n represents the number of classes. ID loss functions have been used in (Huang *et al.*, 2018b; Zheng, Zheng & Yang, 2017d; Wu *et al.*, 2018b; Deng *et al.*, 2018). Generally, this loss is easy to train, and hard samples are automatically mined during the training process. Model overfitting is avoided by integrating label smoothing within the loss function (Zheng *et al.*, 2017d).

Verification loss considers the pairwise relationship between the extracted features of the samples. This is either done with contrastive loss (Varior, Shuai, Lu, Xu & Wang, 2016a) or with binary cross entropy loss where a differential feature generated as a function of two samples under comparison is classified with a classifier as similar or not (Li *et al.*, 2014b).

$$\mathcal{L}_{\text{con}} = \left(1-\delta_{ij}\right)\left\{\max\left(0, \rho - d_{ij}\right)\right\}^2 + \delta_{ij} d_{ij}^2 \tag{1.20}$$

Eqn. 1.20 is the contrastive loss function where d_{ij} is the distance between embeddings and features of two samples, i.e., x_i and x_j. δ_{ij} is the binary label indicator where $\delta_{ij} = 1$ for x_i and x_j having the same identity and $\delta_{ij} = 0$ otherwise.

Triplet loss

When ReID is treated as a retrieval ranking problem, the loss function is designed such that a positive pair or embeddings of similar examples would have a smaller Euclidean distance than

dissimilar ones or negative pairs. Hence, Triplet loss (Section 1.1.3.2) is often used in ReID applications.

1.4 Challenges

Online learning for tracking Be it classical method such as (Wang *et al.*, b) or other deep CNN methods such as (Bhat *et al.*, 2019b; Danelljan *et al.*, 2020; Nam & Han) most of the solutions design a method to learn target appearance online during tracking. This is done to avoid target drift (Nam & Han; Wang *et al.*, b) over time. But as discussed by (Gavves, Tao, Gupta & Smeulders, 2021b), model adaptation by using samples mined by a tracker cannot be fully leveraged due to noise introduced into the samples by a drifting tracker. Hence, it is necessary to have additional mechanisms to detect drifts, classify them, and manage the samples for learning target appearance. In addition to this, sample selection and learning can affect model robustness in terms of being able to track a target. Model adaptation without proper sample variance can cause model drift due to catastrophic forgetting (Bhat *et al.*, 2019b; Gavves *et al.*, 2021b). For example, the target's appearance can be different over time due to changes in viewpoint, and a model that learns the target aggressively can forget the initial appearance of the target. When the viewpoints change, this can once again lead to track drift. This problem has been addressed in the next chapter.

Robust embeddings against computer vision challenges Video Person ReID can be beneficial compared to Image-based ReID depending on the availability of a person tracklet during a query. This can alleviate some challenges in ReID, like occlusion, poor bounding box accuracy of the tackles, and other computer vision challenges, such as lighting conditions and viewpoint changes. Although sequence information has been previously in video person ReID, sequence length has been limited, as shown in several works. One main reason for this was overfitting due to a large number of samples and the blurring effect introduced by average pooling in the temporal dimension. Hence, improving the feature aggregation method while considering video sequences for ReID is necessary. Attention methods can be used to strengthen weighting individual frames, and there is some scope here to use other modalities or additional information,

if any, to help with video feature extraction and leverage the use of lengthy sequences. We address this problem in Chapter 3 of this thesis.

Occlusion in ReID and Tracking Occlusion is another important noise source for trackers and image and video person ReID in online learning. Occlusion can cause feature extractors to focus on regions other than objects, potentially distracting tracking. Occlusion was handled in ReID employing computationally challenging methods. This included using pose estimation Gao *et al.* (2020b), segmentation (Cai, Wang & Cheng, 2019b), and ccosegmentation (Subramaniam, Nambiar & Mittal, 2019a) information. Most of the methods in the literature employ an additional system to estimate body pose or segmentation masks trained on different datasets. Hence, the overall performance depends on several factors, such as the performance of additional systems, noise in the additional datasets, and, most importantly, the computational complexity. Hence, we attempt to solve this problem in Chapter 5 with our proposal that uses guidance and cues from holistic datasets for occluded ReID.

Similar and dissimilar sample pair selection As discussed in the previous sections, sample pair selection is an open problem in contrastive learning or metric learning. Sample selection strategy influences the quality of features learned. At the same time, not all possible pairs can be selected, as this can increase the training complexity and overfit the model on easy samples. This problem has been addressed in Chapter 6 of this thesis.

1.5 Conclusion

Three potential directions have been observed based on the challenges mentioned above and the review of the current literature. One is systematic sample selection and governance of samples maintained for online learning in tracking. The overall design is limited to memory and complexity for real-time systems, making maintaining an optimal sample size of prime importance. Another direction is leveraging video sequences to capture motion information and learning robust embeddings to tackle variations and viewpoint changes. Finally, a third direction

is to focus on occlusion-aware learning of embeddings for ReID and tracking. These directions have been explored in detail in the following chapters.

CHAPTER 2

INCREMENTAL TEMPLATE LEARNING FOR OCCLUSION-AWARE VISUAL OBJECT TRACKING

Madhu Kiran[a] , Eric Granger[a] , Le Thanh Nguyen-Meidine[a] , Rafael Menelau Oliveira Cruz[a] , Louis-Antoine Blais-Morin[b]

[a]Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure,
1100 Notre-Dame St W, Montreal, Quebec H3C 1K3, Canada
[b]Genetec Inc., Montreal, Canada

Paper submitted for publication, Elsevier Expert Systems With Applications, June 2023

Abstract Visual object tracking (VOT) plays a critical role in various video surveillance and monitoring applications, where accurate object localization is essential for tasks such as scene understanding, person re-identification, and face recognition. Deep Siamese trackers have recently gained a lot of attention since they can perform real-time VOT on systems with limited resources, which is crucial in video surveillance. Moreover, adaptive tracking methods have achieved state-of-the-art accuracy, with target samples being collected by the tracker being employed for online learning. However, VOT remains a challenge in real-world applications, particularly occlusions and the changing appearances of objects. This paper proposes VOT as an online learning problem under concept drift. Concept drift detection can be used to trigger updates of the tracker model but, at the same time, can lead to the accumulation of noise in the VOT model. Since VOT models are subjected to different concept (target track) drifts, object appearance changes and occlusion can cause gradual and abrupt concept drifts. In this paper, occlusion-aware drift detection is considered to help differentiate concept drift caused by changing appearance from that caused by occlusion. To mitigate the effects of concept drift, a new adaptive occlusion-aware VOT method is proposed for online incremental learning and memory replay to prevent template corruption. In particular, we propose a mechanism to detect gradual changes in object appearance and dynamically select the corresponding target samples for online adaption. In addition, an entropy-based sample selection strategy is introduced

to maintain a diversified auxiliary buffer for memory replay, alleviating the recurring[1] drift problem. Finally, an occlusion-aware learning strategy is introduced for robust long-term tracking by controlling the influence of occlusion during adaptation. Our proposed method can be integrated into any adaptive VOT model for online learning. Extensive experiments conducted on the challenging OTB-100, LaSOT, UAV123, and TrackingNet datasets indicate the effectiveness of our proposed method and showcase the contribution of its key components, i.e., change detection mechanism, occlusion-aware features and entropy maximization for sampling. Results indicate that integrating our proposed method into state-of-the-art adaptive Siamese trackers can increase the potential benefits of an incremental template adaptation strategy, mitigate the impact of occlusion on VOT, and allow for longer tracking. The code is available: https://github.com/madhukiranets/Adaptive-Siamese-Dimp.git

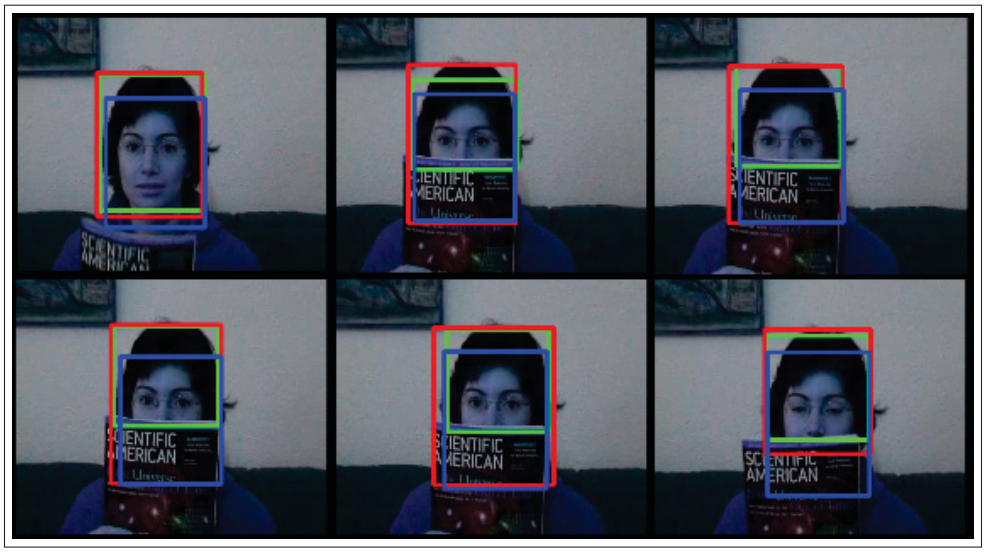

Figure 2.1 Examples of frames from OTB100 dataset (Wu *et al.*, b), showing the ground truth bounding boxes (blue), predictions from our occlusion aware tracker OADiMP (red) and the baseline DiMP (Bhat *et al.*, 2019b) (green). Our proposed method is robust to occlusion compared to the baseline

2.1 Introduction

VOT (or single object tracking, SOT) is a vital processing step in many computer vision applications such as video surveillance. Localizing objects across video frames is often required to continuously track an object and even associate the same objects across different frames to form trajectories. These trajectories are then employed downstream in subsequent processing steps in applications such as person re-identification (Zheng *et al.*, 2016), face recognition (Migneault, Granger & Mokhayeri, 2018), and expression recognition (Gautam & Singh, 2019). The success enjoyed by deep learning (DL) models such as Siamese networks for tracking (Bertinetto *et al.*, 2016; Li *et al.*, a; Dong & Shen, 2018; Zhu *et al.*; Zhang & Peng, 2019; Kiran *et al.*) has led several VOT models to adopt online learning as well (Nam & Han; Danelljan, Bhat, Shahbaz Khan & Felsberg, 2017; Danelljan *et al.*, 2019; Bhat *et al.*, 2019b; Danelljan *et al.*, 2020; Bhat, Danelljan, Van Gool & Timofte, 2020) to improve performance. These trackers can be classified as adaptive tracking methods (or tracking-by-detection methods), and template matching methods (Bertinetto *et al.*, 2016; Zhu *et al.*; Zhang & Peng, 2019; Li *et al.*, a). Adaptive Siamese tracking models differ from template matching methods through their online learning strategy, which typically relies on optimization approaches such as stochastic gradient descent or steepest gradient descent (Bhat *et al.*, 2019b; Danelljan *et al.*, 2020; Bhat *et al.*, 2020; Wang, Zhou, Wang & Li, 2021b). They learn to distinguish an object from its background with an online classifier. In contrast, template matching methods compare an initial object appearance template with a search region and then match the object appearance by interpolating the initial template with the recently generated object template (Valmadre, Bertinetto, Henriques, Vedaldi & Torr, 2017; Zhang, Gonzalez-Garcia, Weijer, Danelljan & Khan, 2019a)

Previously, DL tracker models such TCNN and MDNet (Nam, Baek & Han, 2016; Nam & Han) relied on complex gradient descent methods for online learning. Such online gradient descent methods were not suitable for real-time applications due to their meager frame rate. Recently, trackers such as as (Danelljan *et al.*, 2017, 2019; Bhat *et al.*, 2019b; Danelljan *et al.*, 2020;

[1] The recurring concept is a particular case of concept drift where the concepts already seen in the past reappear as the stream evolves.

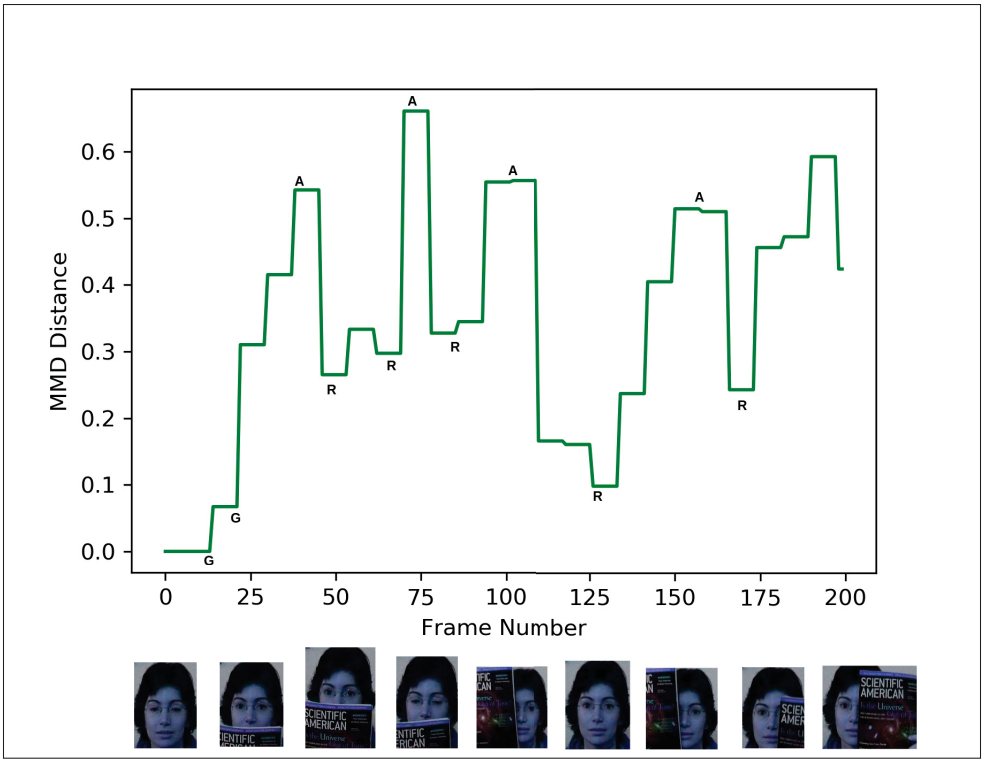

Figure 2.2 Motivation of our proposed method: The plot shows the interference of occlusion during object tracking. It displays the MMD distance between the distribution of object templates around the neighborhood of a given frame and initial object appearance templates. It was produced using a Siamese template-based tracker on the "FaceOcc1" video (OTB dataset). The corresponding object template at a given frame is shown below the X-Axis

Bhat *et al.*, 2020) apply online learning methods using a memory of previously seen templates located by the tracker. The stored memory is then used to optimize a model of the object. Adaptive Siamese trackers generally provide higher accuracy than template matching as they can efficiently update these classifiers.

The following are the main challenges facing object tracking: (1) Appearance change of the object over time, with the change being recurrent or permanent, (2) Online learning of the tracker is affected by using noisy samples localized by the tracker, (3) Efficient model adaptation is necessary to achieve real-time performance, and (4) partial occlusion during tracking can lead to corrupted samples, hampering model updates. The quality of the model deteriorates

significantly with the addition of noise during model update in long-term tracking (Gavves *et al.*, 2021b). It has been shown that model decay due to noise addition is inevitable in long-term tracking (Gavves *et al.*, 2021b).

This paper compares Visual Object Tracking to concept drift during online model adaptation. A concept drift can signify an appearance change or noise in tracking due to issues such as occlusion. Additionally, the localization of the tracker is noisy due to the limitations of the tracker's performance. The localized target then becomes the input to the classifier and learns for the next time frame. Additionally, because the object's appearance changes over time, the distribution of the appearance of the initial target varies, leading to concept drift. Figure 2.2 shows the distribution difference plot of bounding boxes sampled from the OTB dataset (Wu *et al.*, b). Each sample is a representation of deep features using the same backbone as DiMP (Bhat *et al.*, 2019b) tracker. The Y-axis shows the corresponding Maximum Mean Discrepancy (MMD) distance (Rabanser, Günnemann & Lipton, 2019) between features of templates from previous time frames and that of a given time frame to assess the concept drift[2]. During a gradual concept drift relatively small change in appearance occurs over time, often caused by changes in the appearance of the object due to pose variations. Abrupt changes are caused by sudden occlusion or other challenges encountered during tracking. Occlusion that is slowly introduced into the neighborhood of the target location can also cause a gradual drift. These drifts are illustrated in Fig. 2.2, where G stands for gradual and A for abrupt drift. However, the MMD distance in the figure oscillates at some spots. These oscillations are caused by recurring drifts.

Fig. 2.3 displays a plot of the Maximum Mean Discrepancy (MMD) distance of features at a given time frame against the initial template similar to Fig. 2.2, but the plots were generated using two different strategies. The green plot was generated by learning only under concept drift using a mechanism to detect changes or drifts, while the red plot was produced by learning periodically.

[2] We define concept drift as a change in the distribution of data over time in online learning (Agrahari & Singh, 2021) during tracking. It can be observed that there is a correlation between the quality of a given template at a certain time frame and the MMD distance to the initial ground truth templates. Abrupt changes are marked with A, gradual changes with G, and recurring changes with R. The plot shows different types of concept drift or changes that can be observed during tracking, namely gradual, recurrent, and abrupt changes. (detailed in Section 2. C).

It can be seen that the green plot is lower in height, indicating that the current template is located at a shorter distance from the initial template in the feature space. This is because periodically adapting without regard to whether or not a classifier update is really required can lead to corruption, particularly from noise penetrating the tracker model. Hence a gradual change in concept might be a useful situation for the model adaptation, while distractions can cause abrupt changes and must be mitigated. Additionally, recurring changes can be expected, indicating that the object can revert to one of its earlier appearances. A tracker should learn online when a gradual concept drift is detected. Adapting the model or online learning only in the presence of gradual drift reduces the overall complexity by reducing the frequency of learning.

We further analyze occlusion as a common cause of concept drifts during tracking (Gavves *et al.*, 2021b). Fig. 2.1 shows VOT examples from the OTB dataset (Wu *et al.*, b). The green box indicates the output of the baseline DiMP tracker, which does not model occlusion. It can be observed that the tracking performance deteriorates due to the use of partially visible samples for

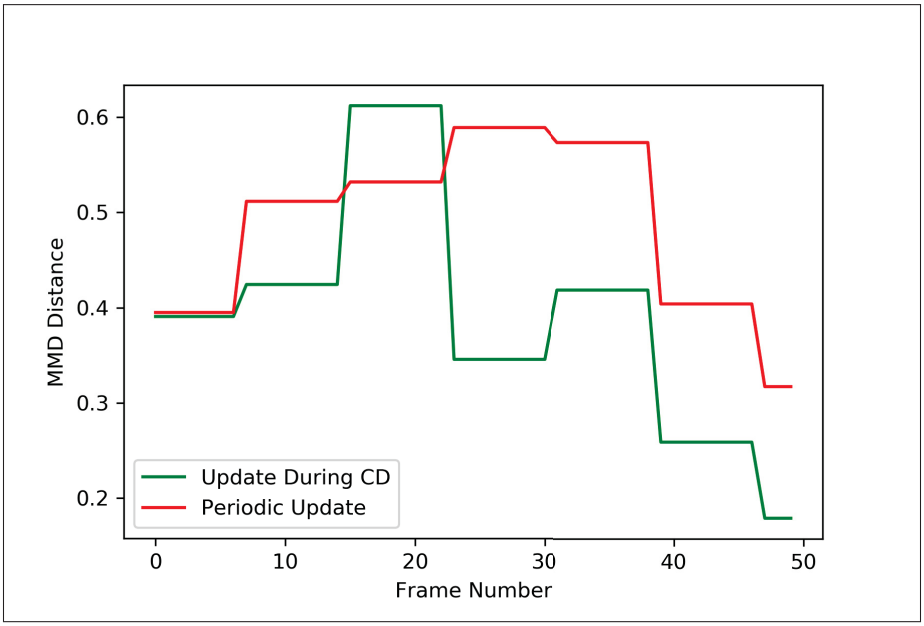

Figure 2.3 Model Corruption by learning under noise The plot shows the MMD distance between learned features and initial feature representation at a given time frame. The red plot was generated by learning periodically, while the green plot was generated by learning when a concept drift was detected

online learning. We also hypothesize that partial occlusion is an important component of noise, which can be integrated into the model during its online model adaptation. Both occlusion appearance changes could cause concept drift. It is, therefore, important to differentiate between concept drifts under different conditions. Fig. 2.1 red box shows the results of our occlusion-aware tracking, which displays comparatively better tracking than the baseline DiMP (Bhat *et al.*, 2019b), and the results are closer to the ground truth bounding box(shown in blue).

Given the challenges faced in tracking in the context of online learning and our hypothesis with respect to it, we propose herein that object tracking be treated as an online learning problem under concept drift. We propose employing a change detection algorithm to classify concept drift during online learning as a gradual and abrupt change. A buffer of samples from past target templates is maintained during tracking memory, and complexity constraints limit the buffer length. When a gradual change in concept is detected, the classifier is trained to adapt to the change in concept drift. During the detection of an abrupt change, no target template adaptation takes place in the temporal neighborhood of the change. We do not propose to react to recurring changes specifically, but the sample selection method for online learning has been designed with the recurring changes in mind, too. Our sample selection strategy ensures the maintenance of samples of high variance in the memory to maintain representation from a variety of object appearances seen. To this end, we extend the entropy maximization-based sampling (Wiewel & Yang, 2021) algorithm to our problem, i.e., online learning under concept drift to maintain high variance samples in a buffer. We extend (Wiewel & Yang, 2021) by proposing a classifier score discretization-based algorithm to alleviate the catastrophic forgetting that often occurs during online tracking by maintaining an auxiliary memory of samples.

A preliminary version of this work was presented in (Kiran *et al.*, 2022), where we mostly focus on online learning with change detection for tracking. We further extend our previous work by considering partial occlusion as one of the main causes of drift and occlusion-aware training to improve target drift during tracking. We introduce a new architecture for occlusion-aware training for object tracking and extensively study occlusion-aware architecture with additional ablations studies in this work. In comparison to our previous work, we further explore change

detection methods to include feature-based change detection methods and show that they are more beneficial compared to the classifier-based change detection method used in (Kiran *et al.*, 2022).

This paper considers learning under a noisy data stream – the input data stream flows from the recursive tracker localization of the object prone to noise. The effect of inevitable model decay due to noisy input data stream (Gavves *et al.*, 2021b) is reduced by training the network end-to-end with occlusion awareness through the provision of artificially occluded samples. Occlusion-aware training helps the features focus on the visible regions and less on the occluded parts, thereby reducing the introduction of noise. The memory buffer size reduction and periodic update rates were previously discussed in (Danelljan *et al.*, 2017). The efficiency of periodic updates is limited by the fixed update rate across a video. Our approach differs in that both memory buffer maintenance and model update are triggered by change detection rather than by a periodic update. Moreover, (Danelljan *et al.*, 2017) improves the memory buffer efficiency by representing the extracted features as a mixture of Gaussian but determining the number of components is complex, which makes our proposed entropy maximization-based method suitable for the proposed deep features-based tracking model. Previously, (Danelljan, Hager, Shahbaz Khan & Felsberg, 2016) reduced the effect of corrupted samples due to occlusion in online tracking by carefully scoring collected samples but did not consider occluded samples exclusively. In contrast, we avoid collecting noisy samples through the occlusion-aware training of backbone features. Model updates are only performed only during gradual change detection. Common problems such as partial occlusion would be classified as abrupt changes due to the occlusion-aware training of the backbone involved with them.

The contributions of this paper are summarized as follows.

- A new method, "OADiMP" is introduced for adaptive occlusion-aware VOT, where target samples collected by the tracker are employed for online learning. It is formulated as an occlusion-aware concept (target track) drift problem, where the VOT model is only updated upon the detection of concept drift.

- A change detection mechanism is proposed to detect concept drift and trigger online adaptation of the occlusion-aware model and address the problem of concept drift in online tracking.
- Given that occlusion and changing object appearance are among the main causes of track drifts, we introduce occlusion-aware training for backbone feature extractors to improve the VOT model's robustness to occlusion.
- An entropy maximization sampling strategy that uses discretized classifier scores is proposed to maintain a diversified auxiliary buffer of templates for memory replay, which alleviates the problem of recurring drift.
- An extensive set of experiments are carried out on OTB-100 (Wu *et al.*, b), LaSOT (Fan *et al.*, 2019), UAV123 (Mueller, Smith & Ghanem, 2016), and TrackingNet (Muller, Bibi, Giancola, Alsubaihi & Ghanem, 2018) datasets. Results indicate that by integrating our proposed method into several state-of-the-art adaptive Siamese trackers, including DiMP (Bhat *et al.*, 2019b), PrDimp (Danelljan *et al.*, 2020), and SuperDiMP (Choi, Lee, Lee & Hauptmann, 2020), one can significantly improve their performance by mitigating the impact of occlusion.

The rest of this paper is organized as follows. Section II provides a critical analysis of the literature on deep Siamese and online incremental tracking and change detection. Section III describes our proposed occlusion-aware VOT based on change detection and incremental template learning. Finally, the experimental methodology and results obtained with the proposed approach are presented and discussed in Section IV.

2.2 Related Work

2.2.1 Siamese Tracking

The Siamese family of trackers evolved from Siamese networks, Pioneered by SINT (Tao, Gavves & Smeulders, 2016) and SiamFC (Bertinetto *et al.*, 2016). They are trained offline for similarity matching with metric learning techniques. A large dataset is required to train these networks so that generic feature representation is learned, allowing them to distinguish the object foreground from the background for tracking. SiamRPN (Li *et al.*, a) further improved

on the pioneering work by employing region proposals to produce a target-specific anchor-based detector. Then, the Siamese trackers that followed mainly involved designing powerful backbones (Zhang & Peng, 2019; Li & Zhang, 2019), or more powerful proposal networks, as in (Fan & Ling, 2019). ATOM (Danelljan *et al.*, 2019) and DIMP (Bhat *et al.*, 2019b) are robust online trackers that differ from the general offline Siamese trackers. Distractor-aware training and domain-specific tracking (Zhu *et al.*; He, Luo, Tian & Zeng) are additional paradigms characterizing Siamese trackers. Recently, (Huang *et al.*, 2021) attempted to model spatiotemporal attention and n (Shen *et al.*, 2022) leveraged unsupervised learning for Siamese tracking.

2.2.2 Online Learning for Tracking

In the paper (Zhong, Bai, Li, Zhang & Fu, 2018), an LSTM is incorporated to capture long-term relationships during tracking, transforming the Visual Object Tracking (VOT) problem into a sequential decision-making process that utilizes reinforcement learning (Duman & Erdem, 2019) to select the best model for tracking. Other approaches, such as (Valmadre *et al.*, 2017) and (Zhu *et al.*), employ online model updating through a moving average-based learning approach, where the target region extracted from the tracker output is integrated into the initial target. Similarly, in (Song *et al.*, 2018c), a generative model is learned using adversarial learning to generate shifted versions of target templates from the original template, and an adversarial function is employed to determine if the generated template belongs to the same distribution. The work in (Yang & Chan, 2018b) utilizes an LSTM to estimate the current template by storing previous templates in a memory bank, while in (Guo *et al.*, 2017), a transformation matrix is computed with regularized linear regression in the Fourier domain based on the initial template. Furthermore, (Yao, Wu, Zhang, Shan & Zuo, 2018) employs online SGD learning to determine the updating coefficients of a correlation filter-based tracker.

These methods update the tracker based on the tracker output as the reference template while building upon the initial template. Kiran et al. propose a generative target update for the UpdateNet architecture that utilizes temporal cues in (Kiran *et al.*, 2022). They also evaluate

tracker confidence online to prevent noisy updates. These approaches differ from our proposed method as they rely on template-matching-based tracking and do not perform online learning. Similarly, the work in (Zhang, Peng, Fu, Li & Hu, 2020b) introduces online model prediction but employs a fast conjugate gradient algorithm for model prediction. They estimate foreground score maps online and combine the classification branch through weighted addition. The template is updated after every frame, following certain sanity checks such as evaluating tracker confidence. The approach in (Dai *et al.*, 2020) is similar to ours in terms of considering tracker updating. However, their approach involves a computationally expensive model that operates at a slow speed (13fps) and employs a trained mechanism to determine when the tracker needs an update.

2.2.2.1 DiMP family of trackers

Online learning for deep trackers (MDNet) was initially introduced by (Nam & Han). They learn a classifier online with gradient descent. Although they show improved performance as compared to other deep trackers, they demonstrate poor tracking speed of less than three fps. (Bhat *et al.*, 2019b; Danelljan *et al.*, 2020) differ from MDNet due to their real-time performance and online learning strategy. They propose a model where an adaptive discriminative model is learned online by the steepest gradient descent method. DiMP family of tracking architectures consists of a target classification branch and a bounding box estimation branch. The tracker uses a discriminative loss to learn to differentiate the object from the background, while a specialized gradient descent technique is used for online learning. The target classification branch takes an input of training images along with annotations for localization to extract deep features with the backbone network.

2.2.3 Online Incremental Learning

The process of incremental learning (Wiewel & Yang, 2021; He, Mao, Shao & Zhu, 2020a) involves learning new classes or distributions of existing classes using sequential data. However, this approach often leads to poor performance on previous tasks due to knowledge corruption or

catastrophic forgetting. To overcome these issues, regularization methods employ constraints by adding a suitable regularization term (Kirkpatrick *et al.*, 2017; Zenke, Poole & Ganguli, 2017), while structural methods freeze weights and expand the architecture along with new tasks (Rusu *et al.*, 2016; Yoon, Yang, Lee & Hwang, 2018). Rehearsal methods, on the other hand, use memory to store previously seen examples to alleviate catastrophic forgetting. For instance, various methods have been proposed to address the issue of catastrophic forgetting in sequential learning scenarios. Gradient Episodic Memory (GEM) is an approach that utilizes examples from previously seen tasks and ensures that the loss on these examples does not increase during the training of new tasks (Lopez-Paz & Ranzato, 2017a). Another approach, known as Generative Replay, employs a Generative Adversarial Network (GAN) to learn and generate samples from previous tasks, which are then mixed with appropriate proportions during the training of new tasks (Shin, Lee, Kim & Kim, 2017). Gradient-based Sample Selection (GSS) maintains a buffer of diverse samples using gradient information, enabling the selection of informative examples for training (Aljundi, Lin, Goujaud & Bengio, 2019b). Furthermore, in work presented in (Wiewel & Yang, 2021), a two-step method is employed to maintain a buffer for mitigating catastrophic forgetting.

These methods contribute to the development of techniques that alleviate catastrophic forgetting by leveraging various mechanisms, such as memory replay, sample selection, and entropy-based approaches.

2.2.4 Change Detection in online learning

Concept drift, which refers to changes in the underlying data distribution, can be classified into four types: gradual, abrupt, incremental, and recurrent. Gradual drift involves a relatively prolonged duration of change compared to abrupt drift. Recurrent drift concepts reappear after some time, while incremental drift involves continuously changing concepts that are also gradual in nature. Drift detection methods can be categorized into four types: sequential, adaptive windowing, fixed cumulative windowing, and statistical. Sequential analysis-based methods predict drift sequentially, such as CUSUM (Gustafsson, 2000) and

Page-Hinckeley (Page, 1954). Window-based methods use a fixed or adaptive window to summarize older and newer information and assess the difference between them, such as ADWIN (Bifet & Gavalda, 2007) and HDDM (Frias-Blanco *et al.*, 2014). Statistical methods employ statistical parameters to predict drift, such as DDM (Gama, Medas, Castillo & Rodrigues, 2004) and RDDM (Barros, Cabral, Gonçalves Jr & Santos, 2017). Some neural network-based models for drift detection include (Lobo, Laña, Del Ser, Bilbao & Kasabov, 2018; Zhang, Chu, Li, Hu & Wu, 2017b), both designed for detecting drift in text data streams. Several drift detection algorithms have been developed for unlabelled data streams, such as margin density for drift detection in (Sethi & Kantardzic, 2017), and k-means clustering and Page-Hinkeley test in (de Andrade Silva, Hruschka & Gama, 2017). It is noteworthy that these methods monitor classifier performance using different metrics but do not make any assumptions about the feature distributions. (Ditzler & Polikar, 2011) proposed using Hellinger distance between the distributions of batches of incoming data and baseline distribution to detect drift. Similarly, (Dasu, Krishnan, Venkatasubramanian & Yi, 2006; Rabanser *et al.*, 2019) find the distribution change between data in the feature space using KL divergence and MMD distance, respectively. These approaches work with the distribution of data without considering classifier performance.

From the general framework discussed above, the following can be summarised. Online learning-based trackers adapt the model based on the recent appearance of an object obtained from tracker output localization. Hence an object during tracking is subjected to various computer vision challenges, including occlusion, as discussed above, leading to tracking drift. Most of the challenges faced by online trackers are managed by the related work by introducing better backbones or feature extractors or improved loss function to enhance discriminative feature learning or by improving online learning strategy. Although the improvements help in overall tracking performance, a few fundamental problems, such as integration of noise from the tracker output into the sample buffers, concept drift during tracking, and drift triggered by occlusion, have been overlooked in the literature. We identify these gaps in the literature and thereby propose concept drift detection and entropy maximization-based model adaptation to tackle the

above-mentioned fundamental problems. Additionally, we propose occlusion-aware training during tracking to enable the tracker to handle occlusion by making the extracted features less sensitive to occlusion.

2.3 Proposed Method

Given the above-discussed challenges in online tracking, we propose to classify concept drift in online learning using a change detection algorithm and model adaptation strategies based on drift classification. Additionally, occlusion is handled during model adaptation and tracking by training the tracker end-to-end with occlusion-aware features to minimize distractions caused by partial occlusion and thereby mitigate the problem of drifting. Occlusion-aware features can also alleviate the problem of model corruption during online learning.

Fig. 2.4 shows the framework of our proposed method fitted with an adaptive Siamese tracker. The proposed tracking framework accepts object templates and utilizes a feature extractor, denoted as Θ, to generate features φ_t (Bhat *et al.*, 2019b; Danelljan *et al.*, 2020; Bhat *et al.*, 2020; Wang *et al.*, 2021b). These features are then stored in a memory buffer that can hold up to $|B|$ instances of object template features from previous frames. The parameter B represents the budget of the memory buffer, and when it is reached, older samples are removed on a first-in, first-out basis. Subsequently, an online classifier is trained using all the instances from the memory buffer by an optimizer, resulting in a model f that is convolved with test image features φ_{S_t}. This process produces a classification score map S, which is used to differentiate the object foreground from the background and achieve localization.

To address the issue of catastrophic forgetting and recurring concepts, we propose using a change detector and training the model only when a concept drift is detected, rather than periodically. Moreover, to alleviate the problem of long-term tracking, an auxiliary memory is used to store older samples with a budget of $|B|$. When the memory is full, older samples are replaced with newer ones based on an entropy maximization algorithm, increasing the overall entropy of the samples in the memory (Wiewel & Yang, 2021). The proposed learning strategy for general

online trackers is summarized as follows.

$$\phi_{t+1} = \arg\min \mathcal{L}\left(x_{1:t}, y_{1:t}\right) \tag{2.1}$$

where ϕ is the parameterised tracker model that minimises the tracker loss $\mathcal{L}$ over the dataset $D = [x_{1:t}, y_{1:t}]$, and x is the image set and y are the tracker predictions at different time frames t. Eqn. 2.1 varies across the DiMP family of trackers see supplementary materials. In (Gavves *et al.*, 2021b), authors have shown that tracker predictions has a noise δ, and the parameter update can be expressed as:

$$\begin{aligned} f_{i,t+1} = f_{i,t} & \underbrace{-2\eta\mathbb{E}\left[\left(f_{i,t} - y_i^*\right) \cdot \left\|\nabla_\phi f_{i,t}\right\|^2\right]}_{\text{true update}} \\ & \underbrace{+2\eta\mathbb{E}\left[\delta_{i,t} \cdot \left\|\nabla_\phi f_{i,t}\right\|^2\right]}_{\text{noisy update}} \end{aligned} \tag{2.2}$$

Eqn. 2.2 has two components: true update and noisy update. Partial occlusion contributes significantly to the noisy update. Hence we propose that an occlusion-aware feature can mitigate the effect of noise δ and propose a new method for learning occlusion-aware features for tracking.

2.3.1 Occlusion-Aware Training

Occlusion in object tracking is often temporary while the object is in motion but poses a serious problem to online tracking, where the tracker slowly learns occlusion and starts drifting. This is often overlooked in the literature for online learning tracking. Thus, in this work, we propose occlusion-aware learning for tracking that learns to attend to and focus on visible regions of the object rather than on occluded regions. In addition to the classification and tracker heads in the architecture, an additional occlusion classification head is proposed. The embedding from the backbone feature extractor network is shared between the tracker head and occlusion classification head. The latter is trained with occlusion-augmented samples to be able to predict a given sample as occluded or not. This step is undertaken during the end-to-end training of the

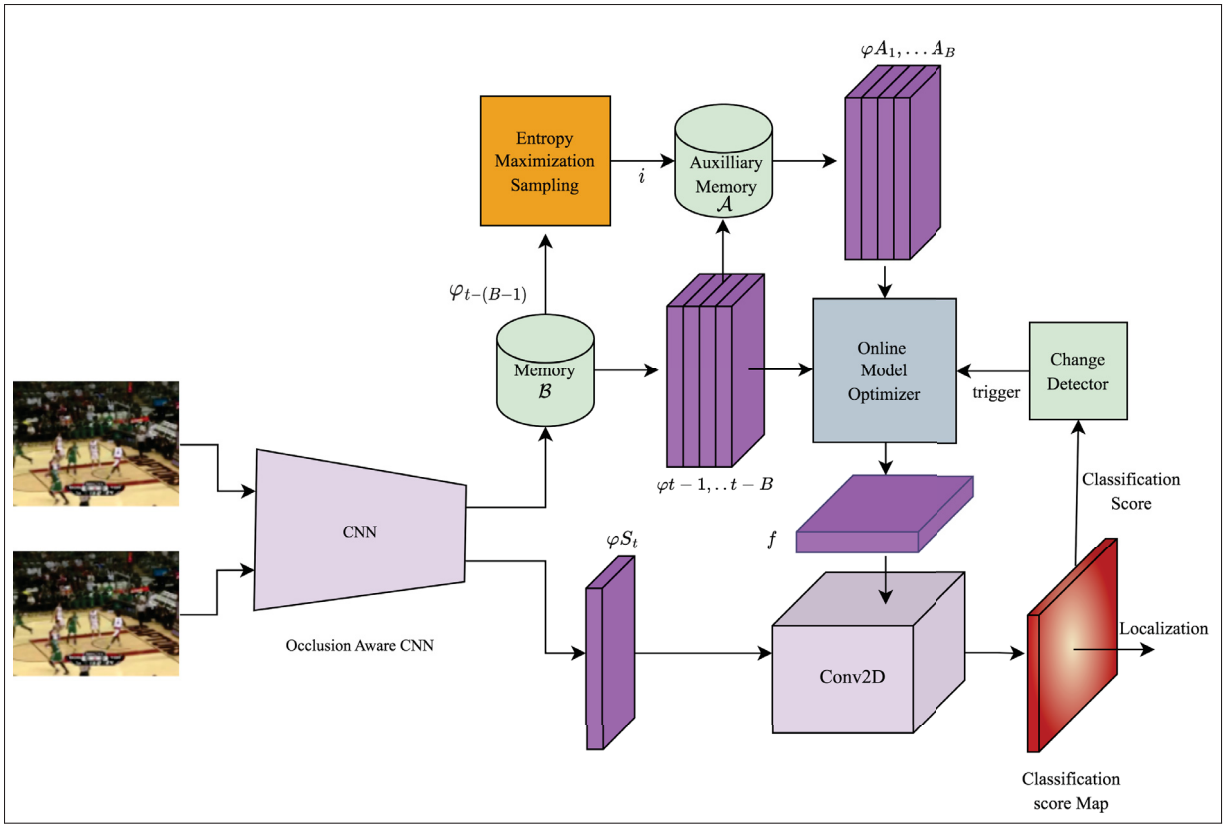

Figure 2.4 Overall framework of our method for occlusion-aware VOT applied to an adaptive Siamese tracker. It relies on a change detection mechanism to trigger adaptation of the occlusion-aware model and additional auxiliary memory to store diverse prior template information

tracker such that the backbone features become robust to partial occlusion, and our proposed model adaptation is then applied to occlusion-aware features during run-time.

Our proposed method can be formulated as an attention model, in which the backbone analyses different samples that are artificially occluded and non-occluded and then uses an occlusion classifier to learn whether or not a given sample is from the occluded sample or not. The joint training with tracking losses and occlusion losses helps integrate expert knowledge (occlusion) into an encoder with prior knowledge(tracking). Since the backbone is shared between the tracker and occlusion classifier, this affects the classifier score for occluded samples and the extracted features from such samples. Hence both our change detection methods, i.e., feature-based and classifier based, respond differently to occluded samples compared to the rest of the samples.

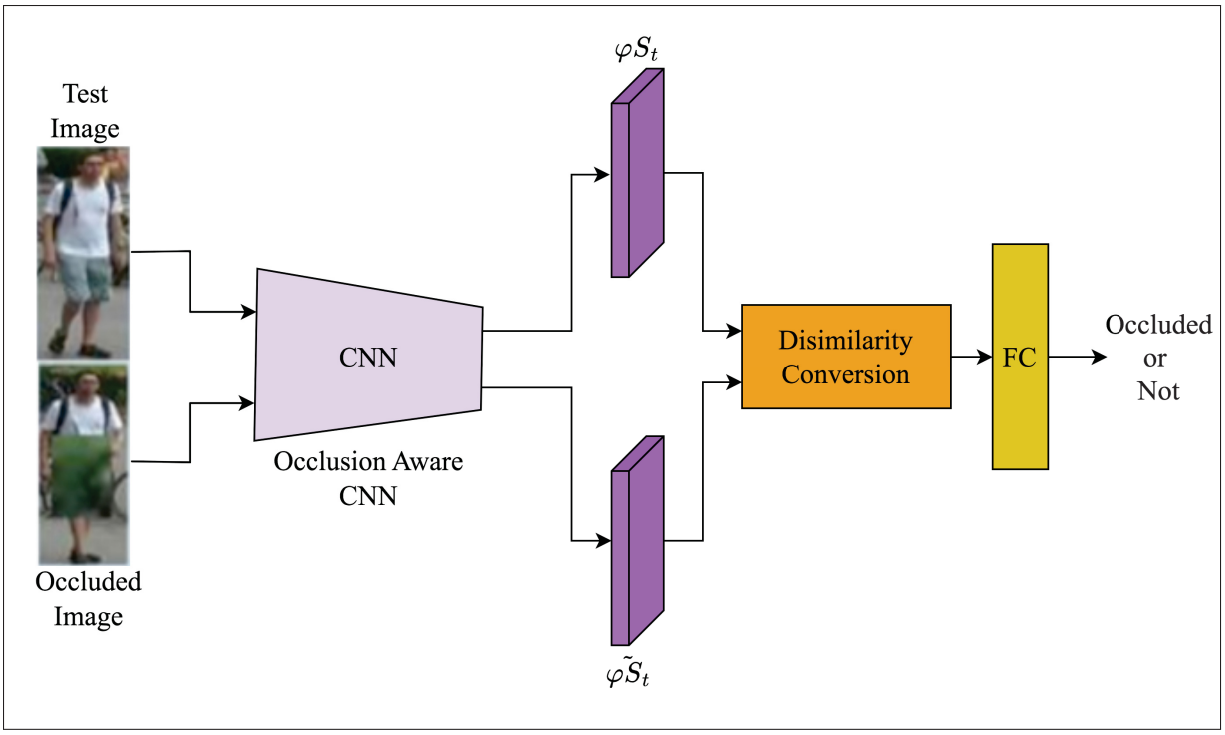

Figure 2.5 Architecture of our proposed occlusion-aware training in the dissimilarity space. Occlusion is simulated on test images, and their features are extracted and converted to dissimilarity space. A 2-class classifier predicts if an input image is occluded

Hence the strong response to occluded samples appears as abrupt changes during change detection when the object being tracked is occluded. We do not adapt the model to abrupt changes and hence avoid learning noisy(occluded sample). Additionally, samples with small occlusions attend more to the visible regions with our occluded aware training, which would act as a mask against occlusion. Fig. 2.5 shows the proposed occlusion-aware training for the backbone feature extractor. An occlusion simulator uses the ground truth location of the object to erase image patches around the object similar to (Zhong, Zheng, Kang, Li & Yang, 2020c). Occlusion-augmented images are labeled 0, and Non-occluded images are labeled 1. Features extracted from original image is φ_{S_t} and that from the occlusion simulator is $\varphi_{\tilde{S}_t}$. R1,2 The given features are converted into dissimilarity space by $\varphi_{dis} = \mid \varphi_{S_t} - \varphi_{\tilde{S}_t} \mid$. φ_{dis} is the representation for the joint features in the dissimilarity space. A fully connected layer is used to classify the given φ_{dis} as occluded (0) or non-occluded (1) based on a given feature being randomly erased or not. The whole tracking system is trained end to end along with the occlusion

classification losses. Tracker is trained to discriminate regions of objects and backgrounds via tracker classification loss. We refer to (Bhat *et al.*, 2019b) for training the classification branch with classification loss. Let $s_\theta(y, x) = \varphi_{S_t}(y)$ be obtained by evaluating the output of the CNN at label y. The classification branch is learned by meta-learning-like settings. Let the threshold T define the target and background region based on the label confidence value z; then tracker classification loss is given by,

$$\mathcal{L}(s, z) = \begin{cases} s - z, & z > T \\ \max(0, s), & z \leq T \end{cases} \tag{2.3}$$

The overall joint optimization relies on the tracker classification and occlusion-aware losses as follows:

$$\mathcal{L}_{joint} = \underbrace{\mathcal{L}(s, z)}_{\text{tracker classfn.}} + \lambda \underbrace{\sum_{c=0}^{C} \left\{O'_i = c\right\} \log \frac{e^{\hat{O}'_i}}{\sum_{c=0}^{C} e^{O\hat{O}_c}}}_{\text{occlusion classfn.}} \tag{2.4}$$

Eqn. 2.4 gives the joint loss for tracker classification and occlusion-aware training. $\hat{y}_i$ is the prediction score of occlusion classifier of the i^{th} example and $c \in \{0, 1\}$ $c = 0$ indicates occluded image and otherwise non-occluded image. λ is the loss balancing parameter and set $\lambda << 0.5$ as tracker classifier training is the most important task among the two.

2.3.2 Change Detection (CD)

Once the backbone features of the tracker are trained end-to-end, we then apply our proposed model adaptation with change detection during the run-time of the tracker to facilitate adaptive online learning. The concept drift problem was previously identified in (Dai *et al.*, 2020), but they did not directly measure or detect concept drift for learning. They proposed memory buffer size reduction and periodic updates of the model anticipating drift. Change detection methods can report a change during concept drifts, and it is often difficult to assess the reason for the change. This can be due to changing appearance of the tracker and also due to occlusion.

Changing appearance (gradual change) needs to be adopted by the tracker, while occluded appearances(abrupt changes) need to be ignored. Occlusion-aware training of the backbone features discussed in the previous section helps differentiate between these changes.

In order to trigger online learning of the classifier, we propose only to adapt the model when a gradual change is detected and abrupt changes are handled by avoiding updates in the neighborhood of the changes. As discussed in Section 2.2.4, this can be achieved either by a classifier-based method where the classifier's performance is analyzed online to detect changes in incoming data distribution or by directly comparing the distribution of incoming data with that of base-line data(data distribution based). Concept drift is a phenomenon that can arise from various sources such as changes in the conditional class probability $P_t(X|y)$, class distributions $P(y)$, or emergence of new classes, as well as changes in the posterior probability $P(c|X)$. In general, concept drift occurs when there is a difference between $P_t(X, y)$ and $P_{t+1}(X, y)$. To address this challenge, the object tracker is first initialized with a ground truth template that represents the object. However, as the appearance of the object may change over time, the tracker must be updated with new object samples collected by the tracker, which may introduce noise into the system (Gama, Žliobaitė, Bifet, Pechenizkiy & Bouchachia, 2014).

2.3.2.1 CD with Classifier

In the proposed framework, a time period $[0, t]$ is considered, where a set of samples $S_{0,t} = d_0, ..., d_t$ is collected, with each instance d_i representing data with features X_i and label y_i, following a distribution $P_t(X, y)$. Concept drift can occur due to changes in conditional class probability $P_t(X|y)$, class distributions $P(y)$, or the appearance of new classes. It is detected when $P_t(X, y) \neq P_{t+1}(X, y)$.

The classification score map, which distinguishes the object foreground from the background for localization, can serve as a good indicator of concept drift. As the distribution of incoming target samples changes, the old model will perform poorly on these samples. The score map is a binary classifier that directly estimates the probability of the predicted label y for a given

object localized by the tracker. Therefore, the change in the classifier score over time is a good estimate of concept change, as it indicates how the old model is performing on new data.

To detect concept drift, an off-the-shelf sequential change detector is employed, which analyses the classification scores as time series data and triggers a flag to indicate a change in classification score. This event triggers the online training of the classifier to obtain a new model φ_{t+1}, which is then used for tracking until the next change is observed.

2.3.2.2 CD with Features

We also propose to employ change detection by distribution distance for detecting drift. From (Ditzler & Polikar, 2011), the baseline data (reference distribution) and incoming data during online learning are presented as batches to the change detection algorithm. Since, in this application, only one sample is available in the incoming data, the data is accumulated to construct a batch of data allowing to extraction of corresponding distributions. Let P denote the baseline distribution and Q that of incoming data, and M the metric to measure the distance between distributions. Let the distance measured at time t be $\delta_H(t)$. The difference between two consecutive distance is $\epsilon(t) = \delta_H(t) - \delta_H(t-1)$. Then a change or drift is reported when,

$$|\epsilon(t)| > \beta(t) \tag{2.5}$$

In Eqn. 2.5 $\beta(t)$ is the adaptive threshold calculated at time t. $\beta(t)$ is computed as follows,

$$\hat{\epsilon} = \frac{1}{t-\lambda-1} \sum_{i=\lambda}^{t-1} |\epsilon(i)| \tag{2.6}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=\lambda}^{t-1} (|\epsilon(i)| - \hat{\epsilon})^2}{t-\lambda-1}} \tag{2.7}$$

In Eqns. 2.6 and 2.7, the mean and standard deviation, λ is the last time stamp during which a drift or change was detected. Finally, the adaptive threshold $\beta(t)$ is obtained below, where γ is the number of standard deviations from the mean, set as three times the standard deviation (Rabanser *et al.*, 2019) empirically.

$$\beta(t) = \hat{\epsilon} + \gamma\hat{\sigma} \tag{2.8}$$

Various metrics can be used to measure the distribution distance like Hellinger distance (Ditzler & Polikar, 2011), KL Divergence (Dasu *et al.*, 2006), and Maximum Mean Discrepancy (MMD) distance (Rabanser *et al.*, 2019). MMD is defined by the idea of representing distances between distributions as distances between mean embeddings of features. Given the recent success of MMD-distance in the measuring domain/distribution shift used in various applications, it was considered in this work.

2.3.3 Entropy Maximization Sampling for Auxiliary Memory

The impact of utilizing a large number of samples to train a classifier can negatively affect the overall runtime of the tracker, rendering it unsuitable for real-time scenarios. To address this issue and ensure that object representations from different time frames are adequately captured, it is crucial to maintain a limited set of samples with maximum variance. To tackle this challenge, we propose the incorporation of an auxiliary memory, denoted as $\mathcal{A}$, which possesses a buffer size of A and is responsible for storing older samples that the main memory has discarded.

By employing the auxiliary memory, we aim to alleviate the problem of catastrophic forgetting that arises due to the limited capacity of the main memory. This auxiliary memory will retain some of the older representations of the object. To facilitate the adaptive Siamese tracking task, which can be viewed as a variation of incremental learning from pseudo labels (labels generated by the tracker for incoming data), we propose a methodology to transform the problem into

binary classification. It is worth noting that the majority of prior research in this domain, as discussed in Section 2, has predominantly focused on multi-class incremental learning.

The process begins by populating the auxiliary memory with samples retrieved from the main memory in a First in, First Out (FIFO) manner. Once the buffer reaches its maximum capacity of A, we employ an entropy maximization algorithm to ensure that samples with high entropy are maintained. In the main memory, each sample is stored as a pair, accompanied by its corresponding classifier score. The classifier scores denoted as S, are discretized and treated as labels. Only samples with classifier scores exceeding a predefined threshold τ are considered for retention.

To adapt the originally designed entropy maximization sampling, which was initially intended for multi-class problems (Wiewel & Yang, 2021), to the context of one-class classification, certain modifications are required. Notably, a crucial distinction between our proposed method and the work described in (Wiewel & Yang, 2021) lies in the fact that the latter pertains to a multi-class continual problem, whereas we are primarily concerned with a single-class incremental learning problem that is susceptible to concept drift. The key challenge in adapting the entropy maximization sampling approach to a single class lies in its reliance on multiple labels, which are readily available in a multi-class scenario, as well as its requirement for a process that maximizes the variance across these labels. To overcome this challenge within the context of a single class, we propose the discretization of classifier confidence scores through binning across the continuous range of these scores. The score is discretized into a range from 0 to Y using a linear mapping function LD, which divides the interval between 0 and 1 and assigns discrete values to the intermediate ranges. As a result, each sample is associated with its corresponding label y. Let $\mathcal{T} = \varphi_j, y_j \mid 1 \leq j \leq B$ represent the samples in the auxiliary memory, where j denotes the index of the sample in the auxiliary memory and y ranges from 0 to Y.

For online model training, we utilize a buffer $\mathcal{B} = \varphi_k, y_k{}_{k=1}^{B}$, which is employed for rehearsal purposes to incorporate previously encountered samples. When an individual input-output pair

(φ_k, y_k) is received, the entropy maximization algorithm determines which old sample should be replaced from the buffer. Both the main and auxiliary buffer images are sampled during runtime to train the model. It is essential for the samples in the auxiliary memory to exhibit the maximum possible diversity in order to avoid catastrophic forgetting. Drawing inspiration from (Wiewel & Yang, 2021), we approach the selection of a sample for rehearsal as a joint distribution $P(\varphi, Y) : \mathbb{R}^S \times \mathbb{Z} \rightarrow [0, 1]$ involving random variables φ and Y. The Shannon entropy of φ and Y is given by $H(\varphi, Y) = -\mathbb{E}[\log P(\varphi, Y)]$. Since Y represents the discretized classifier score for the sample, it can be viewed as a representation of the various appearances of the sample. As $P(\varphi, y)$ becomes more predictable, the entropy tends to approach zero. Our objective is to maximize the diversity of φ, which is akin to the concept of maximum entropy sampling. The Shannon entropy can be reformulated as:

$$\begin{aligned} H(\varphi, Y) &= -\mathbb{E}[\log P(\varphi \mid Y)] - \mathbb{E}[\log P(Y)] \\ &= H(\varphi \mid Y) + H(Y). \end{aligned} \tag{2.9}$$

It can be observed from Eqn 2.9 that the overall entropy depends on $P(Y)$ and $P(\varphi)$. Minimizing $P(Y)$ involves selecting the most common label as the criterion for sample replacement. On the other hand, within samples belonging to the same label, there may be numerous repetitions of similar examples. Minimizing $P(\varphi \mid Y)$ involves estimating the distribution of $P(\varphi \mid Y)$. Similar to (Wiewel & Yang, 2021), we utilize Kernel Density Estimation (KDE) to estimate $P(\varphi \mid Y)$, which is defined as follows:

Algorithm 1 Entropy based sample selection with discrete classifier scores

Require: Target feature sample φ, Classifier Score S, Auxiliary Buffer $\mathcal{A}$, Budget A, **B** main buffer

Ensure: Updated $\mathcal{A}$, BA training samples

if $|\mathcal{A}| < |B|$ **then** ▷ Cardinality check

$y \leftarrow D(S)$ ▷ Discretize classifier score to get a label

$\mathcal{A} \leftarrow \varphi, y$ ▷ Add to the buffer until full

else

$C \leftarrow$ samples of majority label in $\mathcal{A}$

for $\varphi \in C$ **do** ▷ i, j are indices in loop

$d_i \leftarrow \min_{\mathbf{x}_j} \in C \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2$

end for

$i \sim P(i) \leftarrow (1 - d_i) / \sum_j \left(1 - d_j\right)$ ▷ $P(i)$, probability of sample i

$\mathbf{x}_i, y_i \leftarrow \mathbf{x}, y$

$BA \leftarrow \mathcal{B}$+$\mathcal{A}$

end if

$$P(\mathbf{x} \mid y) \approx \frac{1}{M_y} \sum_{k=1}^{M_y} K\left(\mathbf{x} - \mathbf{x}_y[k]\right) \tag{2.10}$$

Here, M_y represents the number of examples with label y, and $K : \mathbb{R}^S \rightarrow [0, 1]$ denotes a Gaussian kernel function. Given these approximations, an algorithm is needed to maximize the joint entropy $H(\varphi \mid Y)$. To begin, we maximize $H(Y)$ by maintaining an equal number of samples with labels Y across all classes. This is achieved by identifying the majority label Y and replacing the sample in the majority class with a probability $(1 - d_i) / \sum_j \left(1 - d_j\right)$, where d_i represents the minimum distance of x_i to all examples of the same class in the buffer. The complete process is detailed in Algorithm 2.

In summary, our proposed method operates as follows: When the auxiliary buffer is not full, all samples exiting from the main memory are added until the auxiliary buffer reaches its capacity.

Once the buffer is full, the majority label is selected based on the relative frequency of the class score discretized label. An example with this label is chosen for replacement, with a higher probability of being replaced if its minimum distance to other examples of the same label in the buffer is small.

During tracking, a batch of samples for training the classifier is created by randomly sampling from the main memory and a fixed set of n samples from the auxiliary memory. The value of n remains constant throughout tracking and is chosen empirically. Additionally, random sampling from the auxiliary buffer ensures that any noisy samples within the buffer do not significantly interfere with the tracking process. The main memory ensures that the current representation is learned, while the auxiliary samples help prevent catastrophic forgetting.

2.4 Results and Discussion

2.4.1 Datasets

The evaluation of our proposed method was conducted on four well-known and challenging Visual Object Tracking (VOT) datasets, including UAV123 (Mueller *et al.*, 2016), OTB-100 (Wu *et al.*, b), LaSOT (Fan *et al.*, 2019), and TrackingNet (Muller *et al.*, 2018). Among these datasets, UAV123 and OTB-100 consist of short tracking videos, while LaSOT and TrackingNet are comprised of large-scale datasets. Details of these datasets are provided below and summarised in Table 2.1.

- **OTB-100 dataset** is widely used as a benchmark in visual object tracking and consists of 100 videos with an average of 590 frames per video. Our reported results were based on the OTB-2015 dataset, which has been widely utilized and is considered to have approached saturation over time (Fu, Liu, Fu & Wang, 2021).
- **UAV123 dataset** comprises 123 low-altitude aerial videos captured from unmanned aerial vehicles (UAVs). This dataset presents various challenges, including small objects, fast motion, and distractor objects.

- **LaSOT** is a high-quality benchmark for large-scale single object tracking, containing 280 test videos. Unlike other datasets, LaSOT videos are longer, with an average of 2500 frames per video. This dataset emphasizes the need for online model adaptation and avoiding catastrophic forgetting.
- **TrackingNet** is a large-scale tracking dataset consisting of videos captured in real-world scenarios. It includes a total of 30,643 videos, with 30,132 training videos and 511 testing videos. The average number of frames per video in this dataset is 470.9.

It is important to note that our proposed method was not evaluated on the VOT datasets (Kristan & et al., 2018; Kristan, 2021), and we did not employ the Precision (Pr) and Expected Average Overlap (EAO) metrics from VOT. The evaluation protocol used in VOT involves tracker re-initialization during tracking failure, which is not suitable for our tracker's intended long-term tracking applications. Hence, we did not include re-initialization during track drift in our tracking protocol.

Table 2.1 Properties of the datasets used for experimental validation

Dataset	Avg. Frames	Classes	Tracking Category
LaSOT	2506	70	Long term
OTB	590	16	Short term
TrackingNet	441	27	Short term
UAV	915	9	Long+Short

2.4.2 Experimental Protocol

Our proposed method is integrated into state-of-art DiMP (Bhat *et al.*, 2019b), PrDiMP (Danelljan *et al.*, 2020), and SuperDiMP (Choi *et al.*, 2020) trackers, all online learning-based trackers. These trackers have been selected for evaluation because DiMP is one of the first efficient real-time online learning trackers, while PrDiMP and SuperDiMP are improvements on DiMP. Both SuperDiMP and PrDiMP trackers are accurate on recent visual object tracking challenge datasets (Kristan, 2021), so they have been selected for our study. We use the pre-trained models provided by the authors and fine-tune them with our proposed occlusion-aware training. We use

Page-Hinckeley (Agrahari & Singh, 2021) test for change detection by classifier approach. This test has been adopted for continuous values such as classifier probability scores. In addition to classifier-based change detection, we evaluate our method using a feature distribution-based change detection algorithm, as previously discussed. We employ the MMD (Rabanser *et al.*, 2019) distance-based metric for measuring the distribution difference between baseline samples and incoming data. We show results for both classifier-based change detection and feature distribution-based change detection. We set the memory buffer to a size of 50, similar to (Bhat *et al.*, 2019b; Danelljan *et al.*, 2020) and the auxiliary memory buffer as 50. The threshold λ for change detection is set to 0.15 empirically by analyzing a set of training videos from the LaSOT dataset by varying the range from 0.05 to 0.2. The buffer size analysis and other hyperparameters are shown as supplementary material.

2.4.3 Integration Into State-of-Art Trackers

Our proposed approach is applied to DiMP, PrDiMP, and SuperDiMP trackers, indicated by OOADiMP, OAPrDiMP, and OAASuperDiMP, respectively, where A refers to adaptive. We compare our method with other template matching-based and adaptive Siamese trackers (Table 2.2). We can observe that our approach significantly improves the AUC accuracy of both state-of-the-art trackers. In particular, accuracy with the UAV and LaSOT datasets is higher than that with the OTB dataset by one percentile because UAV and LaSOT dataset videos are relatively longer than those of OTB-100, and longer videos provide enough samples to improve the quality of samples within the auxiliary memory. SuperDimp and PrDiMP trackers are incremental improvements in ascending order, starting from DiMP. Often the improvements come from the complex backbone and additional neural network weights. But our proposed method can improve the results and perform on par with the next generation of trackers in the DiMP family.

In this paper, trackers are evaluated with the Area Under Curve metric (AUC) of the success plot (Wu *et al.*, b), which measures the overall performance of the tracker under different tracked-to-ground truth overlap acceptance thresholds. The Area Under the Curve was chosen

Table 2.2 AUC accuracy of our proposed method integrated into DiMP50, PrDiMP50 and SuperDiMP on the UAV123, OTB-100, LaSOT and TrackingNet datasets. PH stands for Change detection with the Page-Hinckley test (based on classifier scores), and MMD stands for MMD-distance for change detection with drift detection (based on feature distribution

Tracker	**UAV123**	**OTB**	**TrNet**	**LaSOT**
ECO (Danelljan *et al.*, 2017)	53.2	69.1	-	-
ATOM (Danelljan *et al.*, 2019)	63.2	66.4	70.3	51.5
MDNet (Nam & Han)	-	67.8	-	39.7
SiamRPN++ (Li *et al.*, 2019a)	-	69.6	73.3	49.6
DiMP50 (Bhat *et al.*, 2019b)	65.3	68.4	74	56.9
OADiMP50(Ours)	67.6	69.3	75.7	58.9
PrDiMP50 (Danelljan *et al.*, 2020)	66.7	69.6	75.8	59.8
OAPrDiMP50(Ours)	67.9	71.0	77.1	61.8
SuperDiMP (Choi *et al.*, 2020)	67.1	70.04	78.4	62.6
OASuperDiMP(Ours)	68.2	71.6	79.2	64.4

as the comparison metric as per the standard protocol of One Pass Evaluation (OPE) used for object tracking (Wu *et al.*, b). Our proposed OADiMP, OAPrDiMP, and OASuperDiMP trackers outperform the corresponding baselines. OADiMP improved by 1.6 percentile points over the baseline, and PrDiMP achieved an improvement of 1.7 percentile points on the LaSOT dataset. Similarly, on the TrackingNet dataset, our proposal leads to a 1.5 and 1.2 percentile points improvement, respectively.

Results of our proposed method also include the results obtained with different change detection methods discussed earlier. PH in Tab 2.2 stands for Page-Hinckley test (based on classifier scores), and MMD stands for Maximum Mean Discrepancy (based on drift detection).

Fig. 2.6 shows the comparison of success and precision plots of our proposed method along with state-of-art VOT models. It shows the effectiveness of our approach under different scenarios. From the plots, we achieve a better AUC score in comparison with methods without our proposed change detection and sample selection. PrDiMP and PrDiMP are relative improvements over the

DiMP tracker with significant architectural changes or changes with additional loss functions. However, our proposed method achieves a similar improvement with a simple computationally efficient technique as change detection-based online learning and entropy-based sample selection that can be used to improve the performance of any algorithm from the DiMP family of trackers. The success of our proposed method largely depends upon the length of the input video, as with longer videos, the chances of diverse sample selection are improved. For example, our proposed methods i.e, OADiMP, OAPrDiMP, and OASuperDiMP, improve the performance of respective baseline methods by 2% on the LaSOT dataset which is a long-term tracking dataset.

In addition to improving the tracking accuracy, we also show improve the speed of the trackers. When compared to other methods in the literature, which incrementally improve the accuracy upon common baseline methods often shows a decrease in tracking performance in terms of speed (Dai *et al.*, 2020). (Dai *et al.*, 2020) improve upon ATOM (Danelljan *et al.*, 2019) baseline while the tracking speed is reduced to 13fps.

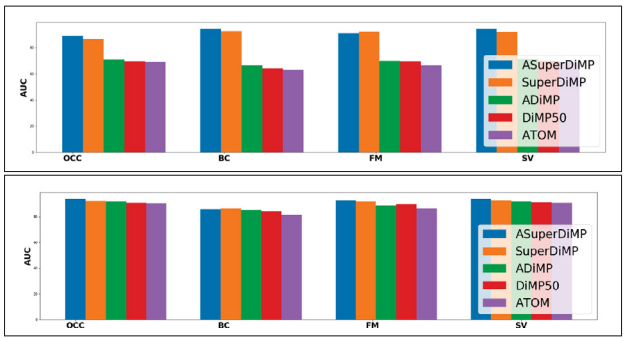

Figure 2.6 AUC scores of success (top) and precision (bottom) plots of VOT models for different video attributes on OTB dataset

2.4.4 Ablation Studies:

In this subsection, we characterize our proposed method using the LaSOT dataset (Fan *et al.*, 2019). We selected the LaSOT dataset for this study due to its longer videos, which are more susceptible to the effects of online adaptation. Our experiments aim to investigate the impact of

Table 2.3 Analysis of different components of our proposed method on LaSOT dataset. The architecture using all the components, including change detection(CD) and Sample replacement with density-based (density) and discrete classifier score-based (class) with occlusion aware features (OCC) achieves an average of 2% improvement on DiMP and SuperDiMP architectures

Tracker	Components					
	CD	**Sample Replacement**			**OCC**	**AUC**
		Random	Density	Class		
DiMP						57.1
	✓	✓				57.9
	✓		✓			58.2
	✓		✓	✓		58.6
	✓		✓	✓	✓	58.9
SuperDiMP						63.1
	✓	✓				62.9
	✓		✓			63.8
	✓		✓	✓		63.9
	✓		✓	✓	✓	64.6

the proposed method's components and key parameters, including the change detection threshold λ.

We perform an ablation study on the proposed method using the LaSOT dataset (Fan *et al.*, 2019), and the results are presented in Table 2.3. The study aims to analyze the effectiveness of different components of the proposed model. The first section of the table examines the impact of the change detector on model updates. We collect samples for the main memory in a First in First Out manner with a buffer size of up to 50 samples. The classifier is updated only when a change in the concept is detected by statistically analyzing the classifier scores for previously seen samples as time-series data. This analysis is performed by the change detector to identify changes in the object's appearance, which signals the optimal time to train the classifier online using the samples from the main memory.

Table 2.4 Analysis of the impact of classifier score-based (PH) and feature distribution(MMD) based change detection methods tested on LaSOT dataset. This experiment was performed with all other components of our proposed system varying the change detection method. (MMD)

Tracker	**AUC**
DiMP baseline	
OADiMP with PH	58.6
OADiMP with MMD	58.9
SuperDiMP baseline	
OASuperDiMP with PH	64.4
OASuperDiMP with MMD	64.4

Table 2.5 Impact of occlusion-aware feature learning evaluated on videos attributed with occlusion in the UAV dataset, evaluated using DiMP and SuperDiMP trackers with all of our proposed components. The experiment analyzed the effect of applying occlusion-aware training in different backbone layers (layers 3,4) and finally on the classification branch (class.). In addition, we analyze the effect of dissimilarity space (Diss.)

Tracker	**Occlusion Aware Training**					**AUC**
	Backbone Layer			**Features**		
	Layer 3	Layer 4	Class	CNN	Diss.	
DiMP	✓			✓		56.7
		✓		✓		57.4
			✓	✓		60.7
			✓		✓	62.8
SuperDiMP	✓			✓		62.4
		✓	✓	✓		62.3
			✓	✓		63.8
					✓	64.9

In this study, we conducted experiments to evaluate the effectiveness of our proposed method on the LaSOT dataset (Fan *et al.*, 2019). The choice of the LaSOT dataset for these experiments

was motivated by its composition of longer videos, which makes it more susceptible to online adaptation. We systematically analyzed the impact of various components and key parameters of our proposed method, including the change detection threshold λ.

To assess the effectiveness of different components in our proposed method, we conducted an ablation study, and the results are presented in Table 2.3. The first part of the table focuses on the influence of the change detector on model updates. Specifically, samples for the main memory were collected using a First-in-First-Out (FIFO) approach, with a maximum buffer size of 50 samples. The classifier was updated only when a change in concept was detected through statistical analysis of classifier scores for previously encountered samples as time-series data. This change detection mechanism allows us to identify instances where the object's appearance is evolving, indicating an opportune moment for online training of the classifier using samples from the main memory.

We conducted experiments wherein the classifier was trained online with randomly selected samples from memory, as denoted by "DiMP with Random" in the table. Additionally, change detection was applied to the classifier's output scores to trigger online training when concept drift was detected. Notably, the change detection-triggered classifier training outperformed all other strategies. This superiority stems from our reliance on the classifier scores to detect concept drift and selectively train the model when necessary. In contrast, other methods employed strategies that did not take into account the classifier's response. Thus, change detection for model updates proves to be an efficient approach in determining the appropriate timing for tracker model updates, as unnecessary updates are susceptible to noise and potential knowledge corruption.

The second part of the table presents the results of experiments conducted to assess the efficacy of our proposed classifier score discretization method in maintaining high variance within the memory. As outlined in the algorithm, our proposed method replaces old samples in memory upon the arrival of new samples. Various criteria determine which samples to replace, including a discrete label based on classifier scores and the density of samples within the same label bin.

We randomly replaced samples from memory upon the arrival of new samples, as indicated by "DiMP and Random sample selection" in the table. Subsequently, we employed density-based replacements considering all samples under a single label, as denoted by "Density." Lastly, we organized samples into individual bins by discretizing classifier scores and assigning discrete values to them. A density-based replacement strategy was then applied to samples within each bin, as indicated by "Class." in the table. Our proposed classifier score-based discretization method yielded improved overall AUC for the tracker, indicating that the method effectively maintains samples in memory with high variance in appearance. This variance ensures that the budgeted memory retains the most informative samples.

2.4.5 Performance of Change Detection Methods:

In this study, we test our proposed system on DiMP and SuperDiMP trackers using classifier score-based and feature distribution-based change detection methods. We use Page Hinckley (PH) test for classifier-based change detection and MMD distance (MMD Drift) based change detection for feature distribution-based change detection. The results are shown in Tab 2.4. It can be observed that the MMD change detection method performs slightly better than PH one considering the DiMP tracker, while for the SuperDiMP one, there is no difference in the performance. The advantage of MMD is that MMD distance is extracted using backbone features, which do not have any influence from previously learned samples with the classifier. Hence is prone to be immune to any noisy classifier update.

2.4.6 Time Complexity:

Table 2.6 displays the frame rate of our tracker versus other VOT models for online learning. The trackers were evaluated on a Linux server with a GTX1080 graphics card. The table shows that our proposed OADiMP, OAPrDiMP, and OASuperDiMP methods run faster than the baselines by an average of 6 and 5 FPS, respectively. This performance improvement is attributed to our novel approach, which differs from the baseline methods by selectively updating the tracker only when concept drift is detected. By doing so, we significantly reduce the frequency of model

updates during the tracking process, leading to a notable reduction in computational complexity. This reduction in computational burden is particularly beneficial in real-time applications, such as video surveillance, where timely processing is crucial to prevent the omission of important events.

Furthermore, our proposed method addresses the challenge of long-term tracking, which is of great importance in various video surveillance applications involving the tracking of objects like vehicles or persons. In such scenarios, it is essential to maintain consistent and distinct identities for tracked objects, even when faced with occlusions or significant appearance changes. Consequently, our method not only improves tracking accuracy but also enhances overall system complexity by effectively handling these challenges.

It is important to note that the memory buffer size represents a bottleneck in our proposed method. With the inclusion of an auxiliary buffer, the total memory consumption increases in proportion to the size of the auxiliary buffer. Thus, careful consideration must be given to balancing memory requirements when deploying our method.

Table 2.6 Average tracking frame rate on LaSOT dataset of state-of-art VOT models that perform online learning. Our method is integrated into DiMP trackers.

Tracker	**Frames/Second (fps)**
ECO (Danelljan *et al.*, 2017)	60
SiamRPN++ (Li *et al.*, 2019a)	35
MDNet (Nam & Han)	3
ATOM (Danelljan *et al.*, 2019)	30
DiMP (Bhat *et al.*, 2019b)	38
OADiMP(ours)	44
PrDiMP (Danelljan *et al.*, 2020)	28
OAPrDiMP(ours)	33
SuperDiMP (Choi *et al.*, 2020)	26
OASuperDiMP(ours)	28

2.5 Conclusion

VOT models that perform online learning are subject to concept drift, limiting their ability to robustly track occluded and changing target object problems to be addressed for robust tracking. We hypothesize that different concept drifts can be observed and classified as gradual, abrupt, and recurring. While a gradual concept drift can help learn changing target appearances, abrupt change often arises as a result of various distractors during tracking, including occlusion and during which model updates must be avoided. Recurring drifts occur when the previous appearance of the object is seen; hence, having a high variance sample buffer for online tracking helps mitigate this problem. A gradual introduction of occlusion in object tracking can cause the tracker to learn the occlusion and start drifting; hence we propose occlusion-aware training to focus on visible regions of the object. Our proposed method is applied to DiMP, PrDiMP, and SuperDiMP trackers, and we show a 2% AUC overall improvement when evaluated on OTB, LaSOT, TrackingNet, and UAV datasets. We also show that our proposed method performs at a higher frame rate as the online classifier training is performed only when a gradual change is detected. A high frame rate is essential for video surveillance applications, and a higher frame rate with our tracker makes it suitable for such applications. For change detection algorithms to be robust, the classifier's confidence must be reliable. In practical scenarios, there is a specific limitation on the reliability of the classifier's confidence scores. In future work, we propose that reliable classifier confidence can be obtained by adopting calibration methods.

CHAPTER 3

FLOW GUIDED MUTUAL-ATTENTION FOR PERSON RE-IDENTIFICATION

Madhu Kiran[a] , Eric Granger[a] , Amran Bhuiyan[a] , Le Thanh Nguyen-Meidine[a] , Louis-Antoine Blais-Morin[b] , Ismail Ben Ayed[a]

[a]Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure,
1100 Notre-Dame St W, Montreal, Quebec H3C 1K3, Canada
[b]Genetec Inc., Montreal, Canada

Paper published in Image and Vision Computing, June 2021

Abstract

Person Re-Identification (ReID) is a challenging problem in many video analytics and surveillance applications, where a person's identity must be associated across a distributed non-overlapping network of cameras. Video-based person ReID has recently gained much interest given the potential for capturing discriminant spatio-temporal information from video clips that is unavailable for image-based ReID. Despite recent advances, deep learning (DL) models for video ReID often fail to leverage this information to improve the robustness of feature representations. In this paper, the motion pattern of a person is explored as an additional cue for ReID. In particular, a flow-guided Mutual Attention network is proposed for fusion of bounding box and optical flow sequences over tracklets using any 2D-CNN backbone, allowing to encode temporal information along with spatial appearance information. Our Mutual Attention network relies on the joint spatial attention between image and optical flow feature maps to activate a common set of salient features. In addition to flow-guided attention, we introduce a method to aggregate features from longer input streams for better video sequence-level representation.

Our extensive experiments on three challenging video ReID datasets indicate that using the proposed approach allows to improve recognition accuracy considerably with respect to conventional gated-attention networks, and state-of-the-art methods for video-based person ReID.

3.1 Introduction

Person Re-Identification (ReID) refers to the problem of associating individuals over a set of non-overlapping camera views. It is a key object recognition tasks, that has recently drawn a significant attention due to its wide range of monitoring and surveillance applications, e.g., multi-camera target tracking, pedestrian tracking in autonomous driving, access control in biometrics, search and retrieval in video surveillance, and human-computer interaction communities. Despite the recent progress with deep learning (DL) models, person re-identification remains a challenging task due to the non-rigid structure of the human body, the variability of capture conditions (e.g., pose, illumination, blur), occlusions, and background clutter.

ReID systems can apply in image-based and video-based settings. State-of-the-art (Ahmed *et al.*, 2015; Bhuiyan *et al.*, 2020; Bhuiyan, Perina & Murino, 2014, 2018; Farenzena, Bazzani, Perina, Murino & Cristani, 2010; Panda, Bhuiyan, Murino & Roy-Chowdhury, 2017; Sun, Zheng, Yang, Tian & Wang, 2018; Quan, Dong, Wu, Zhu & Yang, 2019; Tay, Roy & Yap, 2019) approaches on image-based setting seek to associate still images of individuals captured with a network of non-overlapping cameras. In case of video-based ReID, input video tracklets of an individual are matched against a gallery of tracklet representations, captured with different non-overlapping cameras. A tracklet corresponds to a sequence of bounding boxes that were captured over time for a same person in a camera viewpoint, and are obtained using a person detector and tracker. Compared to image data, video data provides motion information in addition to appearance information which can further enable the system to capture person's body silhouette via post processing. Thus, video-based approaches allow to exploit spatio-temporal information (appearance and motion) for discriminative feature representation.

As illustrated in Figure 3.1, state-of-the-art approaches for video-based person ReID typically learn global features in an end-to-end fashion, through various temporal feature aggregation techniques (Gao & Nevatia, 2018b; Gu *et al.*, 2019; Hou *et al.*, 2019b; McLaughlin *et al.*, 2016a; Subramaniam *et al.*, 2019b). From this figure, the query input to the feature extractor is a video clip (i.e, a set of consecutive bounding boxes extracted from a tracklet) of N frames long. A

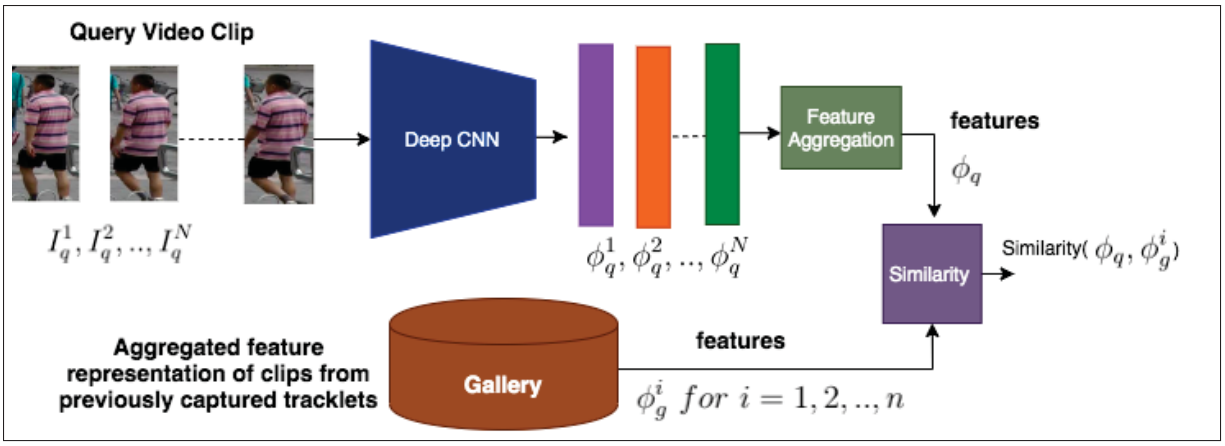

Figure 3.1 Block diagram of a generic DL model specialized for video-based person ReID. Each query video clip from a non-overlapping camera is input to a backbone CNN to produce a set of features embeddings, one per bounding box image. The features are then aggregated to produce an aggregated feature representation each for both video and optical flow clip, which is then matched against clip representations stored in the gallery

single feature vector is extracted by aggregating features from each frame in the query video clip. This is then compared with a gallery containing *n* identities, each one having a set of aggregated feature representations of clips from previously-captured tacklets.

Given a video clip (fixed size set of bounding boxes extracted from a tracklet), the feature extractor (CNN backbone) produces image-level features, while the feature aggregator generates a single feature representation at the clip level, using either average pooling, weighted addition, max pooling, recurrent NNs, etc. (Gao & Nevatia, 2018b) in the temporal domain. Although these aggregation approaches enable to incorporate diverse tracklet information for matching, and can achieve a higher level of accuracy than image-based approaches, they often fail to efficiently capture temporal information which could propagate as salient features throughout the video sequence. Additionally, the performance of state-of-the-art methods decline as the length of video clips grows beyond 4 or 6 frames (Gao & Nevatia, 2018b; Subramaniam *et al.*, 2019b).

Optical flow streams has been previously used to capture the motion dynamics of a person walking in a video stream. Moreover, as shown in Fig. 3.2, the visual appearance of optical flow of a person walking or moving resembles the silhouette of the person, often suppressing

the background static objects. This can potentially be used as a mask on the appearance stream to highlight common saliency between frames. In addition to it highlighting common saliency across frames, the silhouette produced by optical flow can therefore be a good source of spatio-temporal attention.

Fig. 3.2 shows that the flow features are coarse representation of semantic information of moving objects. Unlike action recognition which depends heavily on motion features, accurate ReID is more dependent on appearance features. Hence, there is a scope to combine the strengths of optical flow and appearance (video stream) features for ReID. Previous attempts to include optical flow information into ReID systems (Chung, Tahboub & Delp, 2017; McLaughlin, d. Rincon & Miller, 2016b; opt, 2010) focused on integrating this information just as an additional input stream without exploiting the relationship further between these two streams. This approach has limited effectiveness because optical flow only represents coarse semantic features of moving objects (different from the image stream), and not image-like appearance information. Moreover, the model in (Chung *et al.*, 2017) is related to the two-stream network proposed in (Simonyan & Zisserman, 2014a) that incorporates motion and appearance feature for action recognition. Two-stream networks that are effective for action recognition are less effective for ReID (Chung *et al.*, 2017).

Given the aforementioned justification, we consider the correlation between the visual appearances across motion and appearance streams, along with their individual contribution to motion dynamics. In order to capture long-term spatial information and temporal dynamics in a video clip, a method to aggregate features from longer sequence effectively is presented. This is unexplored in the literature for video person-ReID, thereby undermining the global saliency in the feature representation by using optical flow for both appearance and motion information.

In this paper, a DL model for flow-guided attention is introduced for video-based person ReID that encodes joint spatial attention between features of temporal (optical flow) and spatial (image appearance) information across a tracklet. The proposed Mutual Attention network enables to

Figure 3.2 Example of a sequence of bounding boxes images from the MARS dataset (top row), and its corresponding dense optical flow map (bottom row). The common saliency in the sequence can be observed from the optical flow map

jointly learn a feature embedding that incorporates relevant spatial information from human appearance, along with their motion information, from both appearance and motion streams.

The Mutual Attention network includes both optical flow stream and image stream for ReID, and leverages the mutual appearance and motion information. We also propose a feature aggregation method that allows integrating information aggregation from longer tracklets. Unlike prior work in literature where feature aggregation is achieved by pooling or temporal attention from image feature, the proposed Mutual Attention network relies on a weighted feature addition method over images in a sequence to produce a single feature descriptor based on optical flow and image information. During feature aggregation, a reference frame from each tracklet is selected based on maximum activation from both streams, and weights are assigned for individual features

using image and optical flow information. Attention is enabled from optical flow in both spatial and temporal domain to extract discriminant features for ReID.

The paper is organised as follows. The next section provides some background on DL models for spatio-temporal recognition, optical flow, and attention mechanisms, as they relate to video-based person ReID. Then, Section 3 describes the architecture and associated formulation for our Mutual Attention network. Section 4 and 5 present the experimental methodology and qualitative and quantitative results, respectively. The proposed flow-guided mutual attention network is validated on the challenging MARS, Duke-MTMC, and ILIDS-VID datasets for video-based person ReID. Experimental results show that it can outperform state-of-the-art approaches. Results also indicate the its potential for higher accuracy by processing longer video clips to capture multiple appearance variations.

3.2 Related Work

3.2.1 Image-Based Person-ReID:

The idea of using CNNs for ReID stems from Siamese Network (Bromley *et al.*, 1994), which involves two sub-networks with shared weights, and is suitable for finding the pair-wise similarity between query and reference images. It has first been used in (Yi *et al.*, 2014) that employs three Siamese sub-networks for deep feature learning. Since then many authors focus on designing various DL architectures to learn discriminative feature embedding. Most of these deep-architecture based ReID (Ahmed *et al.*, 2015; Chen *et al.*, 2017a,a; Cheng *et al.*, 2016; Liu *et al.*, 2017a; Varior *et al.*, 2016b) approaches introduce an end-to-end ReID framework, where both feature embedding and metric learning have been investigated as a joint learning problem. In (Ahmed *et al.*, 2015; Varior *et al.*, 2016b), a new layer is proposed to capture the local relationship between two images, which helps modeling pose and viewpoint variations in cross-view pedestrian images. Recent ReID approaches (Su *et al.*, 2017; Zheng, Huang, Lu & Yang, 2017a; Zhao *et al.*, 2017a; Qian *et al.*, 2018; Suh, Wang, Tang, Mei & Lee, 2018a; Zhao *et al.*, 2017b; Saquib Sarfraz, Schumann, Eberle & Stiefelhagen, 2018; Sun *et al.*, 2018;

Bhuiyan *et al.*, 2020) rely on incorporating contextual information into the base deep ReID model, where local and global feature representations are combined to improve accuracy. (Mekhazni, Bhuiyan, Ekladious & Granger, 2020) and (Kiran *et al.*, 2021b) use dissimilarity space instead of feature space for domain adaptation and occluded re-ID. A few attention-based approaches for deep re-ID (Li *et al.*, 2017a; Zhao *et al.*, 2017b; Su *et al.*, 2017) address misalignment challenges by incorporating a regional attention sub-network into a base re-ID model. A thorough review of state-of-the-art on architecture-based approaches underscores the importance of considering local representations, e.g., by dividing the image into soft stripes (Sun *et al.*, 2018) or by pose-based part representation (Su *et al.*, 2017; Suh *et al.*, 2018a; Zheng *et al.*, 2017a; Zhao *et al.*, 2017a; Qian *et al.*, 2018; Zhao *et al.*, 2017b; Saquib Sarfraz *et al.*, 2018). Although these methods have achieved considerable performance improvements, they fail to incorporate temporal information due to their image-based setting.

3.2.2 Video-Based Person-ReID:

Video ReID has recently attracted some interest since temporal information allows dealing with ambiguities such as occlusion and background noise (Gao & Nevatia, 2018b; Gu *et al.*, 2019; Subramaniam *et al.*, 2019b; Hou *et al.*, 2019b; McLaughlin *et al.*, 2016a). An important problem in video-based ReID is the task of aggregating the image level features to obtain one single composite feature or descriptor for a video sequence. (Gao & Nevatia, 2018b) have approached this problem by frame level feature extraction and temporal fusion by using recurrent NNs (RNNs), average pooling, and temporal attention (based on image features). Average Pooling in temporal dimension can be viewed as summing the features of the sequence by giving equal and normalised weights to them. Average pooling of image instance features from a given sequence have proved to be useful in most of the cases, even compared to other DL model based on RNNs or 3D-CNN (Gao & Nevatia, 2018b). 3DCNN has been experimented in (Gao & Nevatia, 2018b; Li *et al.*, 2019c) but have not been very effective in summarising video sequence for reID. But there could be certain case of individual image in a sequence such that they either have

higher noise content or the appearance in the image does not contribute much to an individual's identity, then these become the debatable cases for Average Pooling.

3.2.3 Attention Mechanisms:

Attention can be interpreted as a means of biasing the allocation of available computational resources towards the most informative components of a signal (Hu, Shen & Sun, 2018). A mask guided attention mechanism has been proposed in (Song, Huang, Ouyang & Wang, 2018), where a binary body mask is used in conjunction with the corresponding person image to reduce background clutter. Somewhat similar to (Song *et al.*, 2018), co-segmentation networks have achieved significant improvements in ReID accuracy over the baseline by connecting a new COSAM module between different layers of a deep feature extraction network (Subramaniam *et al.*, 2019b). Co-Segmentation allows extracting common saliency between images, and using this information for both spatial- and channel-wise attention. Other related work for attention in video ReID , (Chen, Lu, Yang & Zhou, 2019) attention is employed in both temporal and and spatial domain. Video stream has been taken advantage of by (Song, Leng, Liu, Hetang & Cai, 2018b) by extracting complementary region based feature by from different frames to obtain informative features as a whole.

3.2.4 Optical Flow as Temporal Stream:

It often serves as a good approximation of the true physical motion projected onto the image plane (opt, 2010). Optical flow has been employed for temporal information fusion in (Chung *et al.*, 2017; McLaughlin *et al.*, 2016b), in a two stream Siamese Network with a weighted cost function to combine the information from both the streams. It uses a CNN that accepts both optical flow and color channels as input, and a recurrent layer to exploit temporal relations. Its important to note that prior to (Chung *et al.*, 2017; Simonyan & Zisserman, 2014a) have used two stream networks but for action recognition. Two stream networks on their own are useful in action recognition as impact of motion cues in action recognition are higher in action recognition than that of ReID (Chung *et al.*, 2017). Therefore here is a necessity to use optical flow in a way

that it can be leveraged for appearance related task. However, traditional two-stream networks are unable to exploit a critical component in re-id i.e appearance across both optical flow and image stream together. Similar to our motivation for using optical flow for appearance along with motion has been discussed in (Ma, Chen, Kira & AlRegib, 2019). Similar to our work, motivation for considering long term temporal relationship has been discussed in (Cho & Foroosh, 2018).

It can be summarised from the above that, as discussed in (Si *et al.*, 2018; Song *et al.*, 2018; Subramaniam *et al.*, 2019b), various saliency feature enhancement methods have been proposed, and they often improve the overall accuracy. Optical flow typically encodes motion information in contrast to appearance information, and hence there is scope to enhancing appearance information from motion and vice-versa. The advantages of long-term information fusion across tracklets are highlighted in (Cho & Foroosh, 2018) and (Ma *et al.*, 2019). Although long-term aggregation of tracklet information can be beneficial, it also suffers from integration of noise while processing longer tracklets. Our proposed approach aims to mitigate this problem through attention-based feature aggregation, thereby taking full advantage of long-term feature aggregation.

3.3 Proposed Mutual Attention Network

We propose a new model for flow-guided attention – the Mutual Attention network. It learns spatial-temporal attention from optical flow thereby focusing on common salient features of a given person during its motion across consecutive frames of a given video clip. Although a two stream Siamese network has been proposed in (Chung *et al.*, 2017), they have included optical flow as an input for re-identification, and do not exploit the full potential of this information. Hence, we seek to leverage the visual appearance of both spatial and temporal streams, i.e., image and optical flow streams, by producing a correlation map between them in the feature space. This correlation map provides attention for both input streams. The temporal information in both the streams is enhanced by enabling the use of longer video and optical flow clips with our proposed feature aggregation method. Our Mutual Attention network therefore includes both

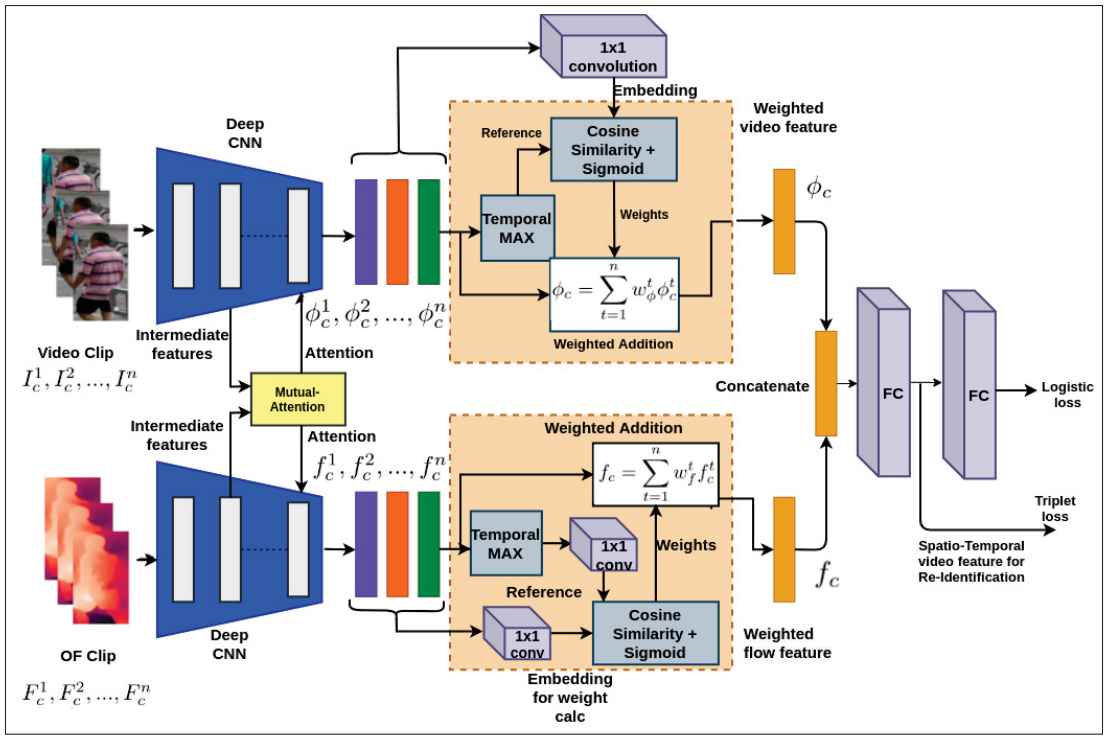

Figure 3.3 Our proposed Mutual Attention network architecture. The system inputs is a video clip and corresponding flow maps. The corresponding features of the inputs attend to each other using our Mutual Attention module. The network outputs a concatenated feature representation from both optical flow and image stream

optical flow and image streams to produce mutual attention, and also to combine the features into an aggregated representation from a video or optical flow tracklet or clip.

As illustrated in Fig. 3.3, the network accepts two streams of input (optical flow and image sequences). At the last layer of the network, the features from the two streams are concatenated after feature aggregation in the temporal domain. While the image stream helps in ReID by focusing on the appearance of the person, optical flow stream helps by capturing motion pattern of a given person. We propose to achieve feature aggregation to produce a feature vector by weighted addition. Our proposed method handles generation of weights that indicate the importance of individual image feature in producing a single video feature leveraging upon mutual attention.

3.3.1 Mutual Attention:

In contrast to the previous method for flow guided attention, we propose to produce cross-stream attention or mutual attention between the optical flow stream and image stream to emphasize areas in the feature space that have high activation across both the streams.

An input video clip is represented by $\mathbf{I_c^1}, \mathbf{I_c^2}, ..., \mathbf{I_c^n}$ and corresponding optical flow estimations $\mathbf{F_c^1}, \mathbf{F_c^2}, ..., \mathbf{F_c^n}$, where c indicates the identity of the video clip of length n, our objective is to extract a discriminative feature vector $\boldsymbol{\phi_c}$ for ReID. Given a video clip and its corresponding flow maps, we extract the features $\boldsymbol{\phi_c}$ and $\mathbf{f_c}$ from the deep CNNs respectively. The expected output is a concatenated feature vector of both optical flow and image features to be used for ReID. Both the CNNs share common architecture but do not share the parameters. Let l be the intermediate layer of the k-layer deep CNN and let appearance CNN be represented by H_{app} and optical flow stream CNN by H_{flow} with a total number of k layers. With $t = 1, 2, 3...n$ we have:

$$\boldsymbol{\phi}_c^t = H_{\text{app}}(\mathbf{I_c^t}) \ , \ \mathbf{f_c^t} = H_{\text{flow}}(\mathbf{F_c^t}) \tag{3.1}$$

If features from layer l are expressed as $\phi^{\mathbf{l}}$, then:

$$\boldsymbol{\phi}_c^{l,t} = H_{\text{app,l}}(\mathbf{I_c^t}) \ , \ \mathbf{f_c^{l,t}} = H_{\text{flow,l}}(\mathbf{F_c^t}) \tag{3.2}$$

Both the features at layer l are of dimensions, $N \times C \times I \times J$, representing sequence length,channels, width, height respectively. The features are then passed through $1x1$ convolution with $Relu$ activation to produce a map of size $N \times 1 \times I \times J$ each. The correlation between the features is given by:

$$\boldsymbol{\rho} = \zeta_{\text{app}}(\boldsymbol{\phi}_c^{l,t}) \odot \zeta_{\text{flow}}(\mathbf{f_c^{l,t}}) \tag{3.3}$$

In the Eq. 3.3, ζ_{app} and ζ_{flow} are the embeddings with 1x1 convolution with $Relu$ discussed above. $\boldsymbol{\rho}$ when activated by a sigmoid function forms the mutual attention map M_c^t between both the streams of input is:

$$\mathbf{M_c^t} = \frac{1}{1 + e^{-\rho}} \tag{3.4}$$

Eqn. 3.3 aims to capture spatial similarities between the Deep CNN features of optical flow and video stream. Optical flow is the result of motion across two frames hence optical flow activation contains information of common saliency between frames. It mostly represents a silhouette of the person. Optical flow could also contain some noisy background motion. Activation from the video stream contains Person specific spatial information along with noises such as occlusion. The noise in both these streams is not common, but the activations in salient regions are common. Multiplying these two activations as in the Eqn. 3.3 identifies similar salient regions in both the streams, thereby reducing the activations due to noise. Therefore, the product fortifies the activations across common saliency in both the streams.

Finally, mutual attention is applied to the intermediate features $\boldsymbol{\phi}_c^{l,t}$ and $\mathbf{f_c^{l,t}}$ at the intermediate layer (by an element-wise multiplication of attention map with feature maps) to obtain mutually attended appearance features $\boldsymbol{\Psi}_{app}$ and $\boldsymbol{\Psi}_{flow}$ to continue feature extraction continues in the remaining layers of the deep CNN to obtain final output features $\boldsymbol{\phi}_c^t$ and $\mathbf{f_c^t}$ for image and flow stream, respectively:

$$\boldsymbol{\Psi}_{app}{}_c^t = \boldsymbol{\phi}_c^{l,t} \odot \mathbf{M_c^t} \ , \ \boldsymbol{\Psi}_{flow}{}_c^t = \mathbf{f_c^{l,t}} \odot \mathbf{M_c^t} \tag{3.5}$$

3.3.2 Weighted Feature Addition:

We propose a method to aggregate image level features to obtain a single feature vector for a given video sequence particularly enabling to use longer video sequences. The appearance features and optical flow features are then concatenated to for ReID during inference, and to learn a classifier during training.

The output from image and optical flow stream CNNs generate $\boldsymbol{\phi_c}$ for a sequence c from instances $\boldsymbol{\phi_c^1, \phi_c^2, ..., \phi_c^n}$ and $\mathbf{f_c}$ from $\mathbf{f_c^1}, \mathbf{f_c^2}, ..., \mathbf{f_c^n}$. The first task is to identify salient feature from a given sequence of features. In our case, a salient feature can be defined as the one that has maximum activation in both image and flow stream. Since the features have been attended by mutual attention, given a sequence, a max operation in the temporal domain for each of the sequence will identify the salient feature among the sequence. We hereafter will refer to this

salient feature as reference frame denoted by $\boldsymbol{\phi_c^{max}}$ and $\mathbf{f_c^{max}}$. In the next step an adaptive weight is generated for each of the features in the sequence based on how close each feature is with the reference feature. This is achieved by applying a cosine similarity between the reference feature and rest of the features int he sequence. The cosine similarity function is not applied directly on the features $\boldsymbol{\phi_c^n}$ and $\mathbf{f_c^n}$. Instead a tiny embedding $\boldsymbol{\epsilon(.)}$ is applied on the $\boldsymbol{\phi_c^n}$,$\boldsymbol{f_c^n}$ and reference feature $\boldsymbol{\phi_c^{max}}$, $\mathbf{f_c^{max}}$ to obtain embeddings $\boldsymbol{\phi_\epsilon^n}$, $\mathbf{f_\epsilon^n}$, $\boldsymbol{\phi_\epsilon^{max}}$ and $\mathbf{f_\epsilon^{max}}$:

$$w_{app}^{n} = \exp\left(\frac{\boldsymbol{\phi_\epsilon^n} \cdot \boldsymbol{\phi_\epsilon^{max}}}{||\boldsymbol{\phi_\epsilon^n}|\,|\boldsymbol{\phi_\epsilon^{max}}|\,|}\right) \tag{3.6}$$

$$w_{flow}^{n} = \exp\left(\frac{\mathbf{f_\epsilon^n} \cdot \mathbf{f_\epsilon^{max}}}{||\mathbf{f_\epsilon^n}|\,|\mathbf{f_\epsilon^{max}}|\,|}\right) \tag{3.7}$$

$$\boldsymbol{\phi_c} = \sum_{t=1}^{n} w_{app^n} \boldsymbol{\phi_c^n} \tag{3.8}$$

A tracklet is a collection of person bounding boxes with same identity. Applying an average pooling operation on an embedding (Gao & Nevatia, 2018b) appears like a sensible solution to summarize tracklet features. However, it is possible that one or more boxes in a tracklet could be occluded, or may contain other kinds of noise that may corrupt an average pooled tracklet. A tracklet can be compared to a cluster since a tracklet is a collection of same identities. Hence, applying the "max" operation on the tracklet is comparable to identifying the densest sample in a given cluster. The embedding that has maximum activation among all other embeddings can be assumed to have the most influence on the cluster center. Hence the Max embedding is used as a reference to calculate similarity with other embeddings in Eq. 3.6, allowing to calculate weights for each of embedding. These weights help determining the closeness of each embedding with the Max embedding (similar to a cluster center) which is the representative of a tracklet. Noisy embeddings that have lower activations can be disregarded by lower weights. The exponential operator ensures non-linearity, as well as assuring a minimum weight of 1 to each of the embedding. Thereby considering all the embeddings in a tracklet for the final representation without completely disregarding noisy features.

A similar approach is followed to accumulate optical flow embeddings with Eq. 3.7. In order to obtain one single embedding for a given tracklet, feature representations in the tracklet are aggregated by weighted addition assigning weights obtained in Eqs. 3.6 and 3.7 to each of the embeddings in a given tracklet. These weights can also be compared to attention-based weights where most important frames in a tracklet are given higher weights, and giving lower weights to noisy frames. From Eq. 3.8 we obtain outputs $\boldsymbol{\phi_c}$ and when used with w_{flow} to obtain $\mathbf{f_c}$ which are aggregated video features for image and flow respectively. These two features are concatenated to form $\boldsymbol{\phi_{cat}}$ which is passed though a fully connected layer of size same as $\boldsymbol{\phi_c}$ to produce the final feature for classification or re-identification.

The network is trained on logistic loss and Triplet loss similar to the method used in (Subramaniam *et al.*, 2019b). During testing the fully connected layer is removed and the remainder of the network is used for feature extraction purpose and to match against those in gallery.

3.4 Experimental Methodology

3.4.1 Datasets:

Experiment are performed on 3 challenging and widely-used datasets for video-based person reID. MARS (Zheng *et al.*, 2016) dataset is one of the largest datasets for video Person-ReID. The dataset has been collected from six cameras with a total of 1261 identities. Another dataset commonly used in literature for evaluating video person ReID is Duke-MTMC (Wu *et al.*, 2018b), (Ristani, Solera, Zou, Cucchiara & Tomasi, 2016c) dataset containing 702 identities and more than 2000 sequences for testing and training each. Duke-MTMC dataset are of higher resolution compared to that of MARS. We also evaluate on ILIDS-VID (Wang, Gong, Zhu & Wang, 2014) dataset, which has a total of 300 identities with videos across two cameras. The dataset is comprised of sequences from two disjoint camera views. One interesting point to be noted with ILIDS datasets is that the tracklets have been generated by hand annotation unlike detector based annotation in MARS dataset. This makes the bounding boxes well aligned in ILIDS-VID dataset enabling the optical flow estimation to be less noisy.

3.4.2 Settings:

We follow the overall system architecture in (Gao & Nevatia, 2018b) (baseline) and (Subramaniam *et al.*, 2019b) (COSAM). They achieved state-of-the-art results on several ReID datasets using the ResNet50 CNN for feature extraction. We propose to use ResNet50 and SE-ResNet50 as our backbone networks to learn features invariant to cluttered background by attending with saliency map obtained from optical flow estimations. We have shown results with each of the networks as backbone separately. The networks have been pre-trained on the ImageNet (Deng *et al.*, 2009) dataset. We experiment at different layers of the ResNet50 to select the ideal location in the network to generate maximum attention with optical flow. To extract video level feature from instance level features, we compare our proposed weighted addition method with that of temporal Average Pooling (AP) and Temporal Attention (TA) based method as illustrated in (Gao & Nevatia, 2018b) and (Subramaniam *et al.*, 2019b). The Shallow CNN in our experimental setup for gated attention is based on AlexNet and the sub-Network for weighted addition is a two layer MLP of size 2048 nodes in each layer.

Common data augmentation methods such as random flips and random crops are followed during training. We use ADAM optimizer to train our model with a batch size of 32. We use a sequence length of 4 to train our model. The flow guided attention has been applied at layer 4 of the ResNet50 and an empirical study of attention at different layers has been presented in the next section. Hence the training setting have mostly been kept similar to our baselines (Gao & Nevatia, 2018b). In order to explore the advantage of attention with optical flow, we propose a separate experimental setup with simple gated attention mechanism i.e using optical flow to attend to image stream alone. This one stream approach, where only image features are used for classification while optical flow is just used as an attention mechanism as described below.

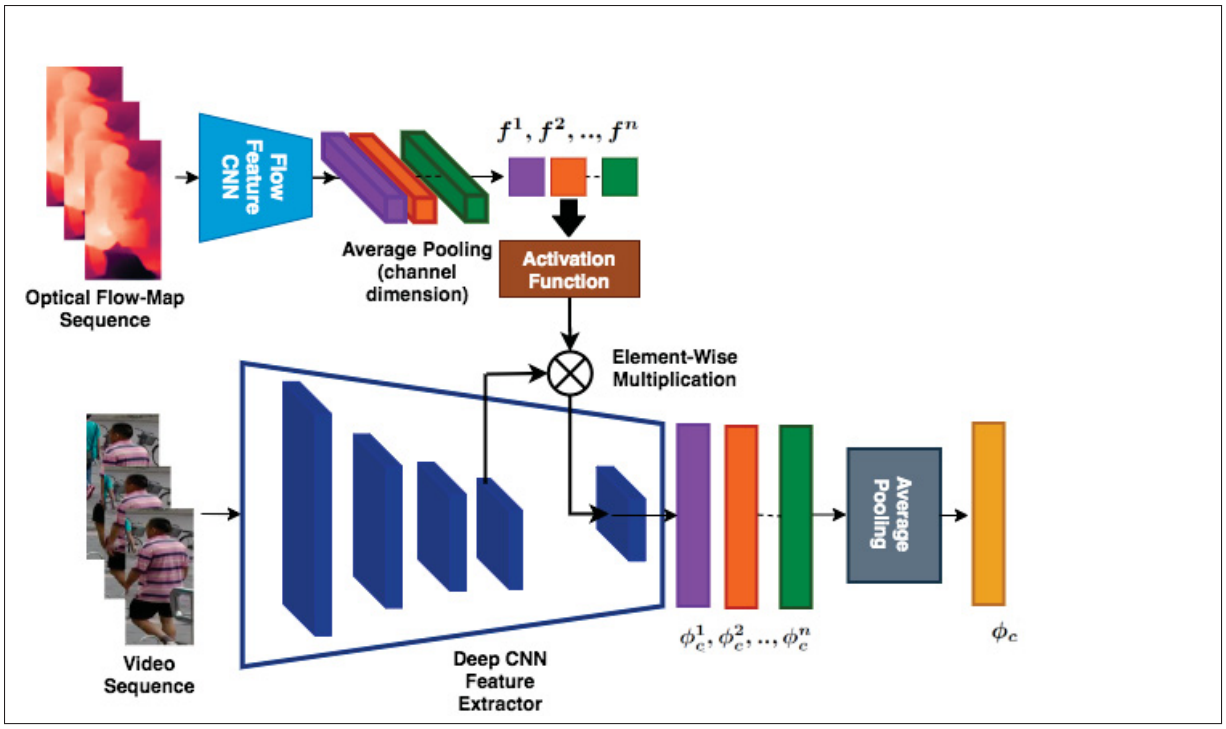

Figure 3.4 Experimental setup for our baseline gated attention network. The system inputs a sequence of bounding boxed images and the corresponding optical flow maps from a given video clip. The features extracted from optical flow are pooled in channel dimension, and multiplied with intermediate layers of deep feature extraction after activation to obtain attended features for ReID. The network outputs a feature vector ϕ_c per video clip

3.4.2.1 Optical Flow Estimation:

To estimate optical flow maps for a given sequence, the LiteFlowNet (Hui, Tang & Loy, 2018) model has been chosen as it is computationally efficient compared to other DL models, while obtaining state of the art performance. We have used the official implementation from the authors to produce flow maps for MARS, Duke-MTMC, and ILIDS-VID datasets. Hence for a given sequence, we input pairs of image I^{t-1}, I^t as input to the LiteFlowNet model to produce flow map F^{t-1}.

3.4.2.2 Baseline for Mutual Attention:

Given input images, $\mathbf{I_c^1}, \mathbf{I_c^2}, ..., \mathbf{I_c^n}$ and $\mathbf{F_c^1}, \mathbf{F_c^2}, ..., \mathbf{F_c^n}$ flow maps for images of a sequence, we extract the features $\boldsymbol{\phi_c} = (\boldsymbol{\phi_c^1}, \boldsymbol{\phi_c^2}, ..., \boldsymbol{\phi_c^n})$ and $\mathbf{f_c} = (\mathbf{f_c^1}, \mathbf{f_c^2}, ..., \mathbf{f_c^n})$ from the deep CNN and the Flow CNN respectively. A shallow CNN (Flow CNN) has been used to extract features from optical flow maps. We rely on shallow CNN to retain spatial coherence in the flow features (see Fig. 3.4).

Let l be the intermediate layer of the k layer deep CNN and let the deep CNN be represented by H with a total number of k layers. Let S denote the shallow flow feature extractor CNN with $t = 1, 2, 3...n$ then:

$$\boldsymbol{\phi_c^t} = \mathbf{H}(\mathbf{I_c^t}) \ , \ \mathbf{f_c^t} = S(\mathbf{F_c^t}) \tag{3.9}$$

If features from layer l are expressed as ψ, then:

$$\boldsymbol{\psi_c^t} = H_l(\mathbf{I_c^t}) \tag{3.10}$$

The features from Eq. 3.10 are then pooled in the channel dimension to produce $N \times 1 \times I \times J$ feature for attention in the spatial dimension. The features are then activated by an activation function to produce a spatial soft attention map($A(f_c^t)$). where A is the sigmoid activation function, and a_c^t is the output of the activation function. Finally attention is applied to the intermediate features ψ at an intermediate layer to obtain activated features Ψ by element-wise multiplication with the activation a_c^t.

$$\boldsymbol{\Psi_c^t} = \boldsymbol{\psi_c^t} \odot \mathbf{a_c^t} \tag{3.11}$$

After gated attention of features at the intermediate CNN layers, the feature extraction process is continued with the rest of the CNN layers. In Eq. 3.12 l, k represents layers between intermediate layer l and last layer k. Then, the output ϕ_c^t of feature extraction is given by,

$$\boldsymbol{\phi_c^t} = H_{l,k}(\boldsymbol{\Psi_c^t}) \tag{3.12}$$

ϕ_c^t is further average pooled in temporal dimension to produce CNN features for re-identification. The network is trained with classification layer similar to our Mutual Attention Network. The gated attention network described here served as one of the baselines for flow guided attention.

3.4.3 Evaluation Measures:

During the training phase we learn the ReID task by training a classifier with identity labels from the single feature extracted for a sequence. The feature tractor produces a 2048×1 size feature vector per sequence. This is the input to train the ReID classifier. During the testing phase, we use the 2048 dimensional feature to measure distance between the test sequence and the sequence from the gallery. We use the Cumulative Matching Characteristic (CMC) and Mean Average Precision(mAP) to evaluate the performance. CMC represents the matching characteristics of the first n query results.

3.5 Results and Discussion

First, this section provides an ablation study the assess the contribution of different modules to the performance of our proposed model. This study is performed on a systems using the baseline ResNet50 CNN architecture (Gao & Nevatia, 2018b), using temporal pooling for feature aggregation. We also include an empirical study that allows selecting the deep feature extraction layer for embedding our attention mechanism, as well as the sequence length. The overall performance is compared against the baselines and state-of-the-art methods. We compare the overall performance on MARS, Duke-MTMC, and ILIDS-Vid dataset.

3.5.1 Flow Guided Attention Fusion:

The first part of our work consists of flow guided attention on the intermediate layer of the Deep CNN used for feature extraction in ReID. Our proposed network consists of flow guided attention on an intermediate layer of the Deep CNN used for feature extraction. Different layers in the Deep CNN have different abstraction level for salient features of the input person image.

Table 3.1 Accuracy of our Baseline (ResNet50) with Temporal Pooling (TP) and our Baseline with Mutual Attention (MA) + TP on MARS dataset, over different layers of ResNet50

Method	maP	Rank-1
Baseline (Gao & Nevatia, 2018b)	75.8	83.1
Baseline+MA (Layer2)	78.2	84.5
Baseline+MA (Layer3)	78.8	84.9
Baseline+MA (Layer4)	**80.0**	**86.6**
Baseline+MA (Layer5)	78.1	84.3
Baseline+MA (All)	51.2	66.1

Hence an experiment was conducted by fusing the flow guided attention at different layers of the Depp CNN applied on the baseline (Gao & Nevatia, 2018b) ReID system, and evaluated on MARS dataset. From the Tab. 3.1 we conclude that the best performance was achieved by attending at layer 4 of the ResNet50 network. This is justifiable from the fact that earlier layers have different abstraction level for salient features – the level of abstraction increases in the later layers but, in the last layer, the spatial coherence is lower than in previous layers. We also conduct an experiment where we apply our MA to multiple layers of ResNet50 (namely to layers 2, 3, 4 and 5), and obtained the results shown in Tab. 1. It can be observed that multi-layer attention provides much lower accuracy than expected. This may be due to the attention from optical flow being sparse, and thus applying them on all layers drastically reduces the overall number of neurons activated across the CNN.

3.5.2 Contribution of Different Modules to the Baseline:

In this subsection we compare Mutual Attention methods to the baseline (Gao & Nevatia, 2018b) since we follow similar overall system architecture. In Tab. 3.2 we compare the baseline and ours with different feature aggregation methods like Average Pooling, Temporal Attention and our weighted feature addition method described earlier. We also compare our Mutual Attention method with single stream with Gated Attention using optical flow described in the

introduction to Experiments section. We can see that just Gated Attention on its own on the baseline (Gao & Nevatia, 2018b) has improved the performances by a large margin on MARS datasets. Our Mutual Attention method further improved the results compared to Gated Attention showing the potential of Mutual Attention between both image and optical flow features. Out Feature addition method used with Mutual Attention improve the results for feature aggregation by a larger margin compared to both Average Pooling and aggregation method from Gated Attention. Tab. 3.2 has additional results with an ablation study for using shared backbone parameters for the dual stream. It can be observed from the table that using shared parameters ("shared param" in table) between optical flow network and ResNet produces much lower results than using separate parameters ("separate param" in table). This is because of the deference in the representation space of both optical flow and image. Although optical flow resembles a

Table 3.2 An ablation study of contribution of different module, i.e., our Gated Attention and our Mutual Attention on the baselines models, using the MARS dataset

Method	**Feature Aggregation**	**mAP**	**Rank 1**
Baseline (Gao & Nevatia, 2018b) (One Stream)	Pooling	75.8	83.1
No Attention (Two Stream)	Pooling	76.7	84.3
Gated Attention (One Stream)	Pooling	77.4	84.6
Mutual Attention (Two Stream)	Pooling	79.1	85.4
Baseline (Gao & Nevatia, 2018b)	RNN	75.7	82.9
Baseline (Gao & Nevatia, 2018b)	Temporal Attention	76.7	83.3
Mutual Attention (Mutual Attention) Shared param	Weighted Addition	76.8	84.1
Mutual Attention (Mutual Attention) Separate param	Weighted Addition	**80.0**	**86.6**

silhouette of the person image, it lacks many fine grained spatial features such as intricate shapes and textures that can be seen in the image space. CNN backbone filters of image based stream tend to learn the different shape and texture information which might not be suitable to learn optical flow information and vice versa. Thereby using separate CNN backbones would enable each of the backbones to learn ideal filters for the corresponding input streams.

Table 3.3 Accuracy of our proposed Mutual Attention (MA), and baselines with no attention, with average pooling, and with RNN for different video sequence length on MARS dataset.

No of frames per sequence	**Baseline (Gao & Nevatia, 2018b)**		**RNN Aggregation**		**Ours**			
	Average Pooling No Attention				**Gated Attention Weighted Addition**		**Mutual Attention Weighted Addition**	
	MAP	Rank 1	mAP	Rank 1	mAP	Rank 1	mAP	Rank 1
2	71.0	81.8	-	-	77.0	84.0	74.8	82.4
4	75.1	83.2	75.7	82.9	77.8	84.8	77.7	85.4
6	74.4	82.7	-	-	77.6	84.5	79.2	85.8
8	73.3	82.0	76.2	82.5	77.3	84.2	79.3	86.4
16	-	-			72.5	82.9	**80.0**	**86.6**

3.5.3 Effect of Sequence Length:

The length of the sequence has an effect on the representative power of final aggregated feature. This in-turn influences the performance of various feature aggregation methods. Therefore in this subsection we analyse the effect of sequence length on different feature aggregation methods such as Temporal Pooling, Flow guided weighted Addition applied on our method. Hence in Tab. 3.3 we have shown results of flow guided attention on ResNet50 architecture with both the feature aggregation methods. It can be seen that at the sequence length of 4 we obtain ideal results for most methods in the literature as well as for the simple gated attention method. But our Mutual Attention method demonstrate the ability to aggregate additional features and hence we could use a sequence of length 16 with Mutual Attention. This is a crucial result as we demonstrate ability to aggregate additional features and keep improving results until a sequence length of 16. Longer sequences have attributed to long term better motion and appearance

features. At the same time in other methods in literature, simple averaging adds additional noise to the features with longer sequences. Our weighted addition methods weights the individual feature based on importance and relevance thereby reducing noise with longer sequences.

3.5.4 Comparison with State-of-the-Art:

We report the performance of our method with backbones ResNet50 and SE-ResNet50 (Hu, Shen & Sun, 2018b) separately with our Mutual Attention and weighted feature aggregation method on MARS , Duke-MTMC datasets and ILIDS-VID (Wang *et al.*, 2014) in the Tab. 3.4 and Tab. 3.5 compared with related works. As mentioned earlier we have selected (Gao & Nevatia, 2018b) as our baseline. It can be observed that from our baseline. we have improved by a large margin on both mAP and Rank1 metric. Our method has also outperformed most of the state-f-the-art methods including some of the best existing methods. We have also shown the advantage of our method compared to other optical flow based methods (Zhang *et al.*, 2019b; Wu *et al.*, 2019b; Chen *et al.*, 2018a). Although (Liu, Wu, Wang & Chien, 2019a) have demonstrated State-Of-the-Art results, we do not compare with them as their evaluation strategy is different from that of the commonly followed method in literature. We attribute our performance gain compared to the baseline on both flow guided attention and our feature aggregation technique. It can also be observed that from our proposal Mutual Attention method performs best demonstrating that optical flow and image stream can attend to the salient regions of each other. Our proposed Mutual Attention method could be integrated with any back-end architecture from some of the strong baselines to achieve better performance. Since our main aim was to evaluate the contribution of Mutual Attention clearly, we choose a simple baseline (Gao & Nevatia, 2018b). Compared to the results on other datasets, the margin of improvement is higher with ILIDS-VID dataset due uniformly centered hand-annotated person cutouts of ILIDS-VID dataset. Results also suggest that our proposed approach is vulnerable to quality of person detection and tracking.

We compare the complexity and attention method used in recent SOA methods along with their complexities in GFLOPs. It can be observed that the complexity of our method is somewhat

Table 3.4 Accuracy of our proposed vs state-of-the-art methods evaluated on the MARS and Duke-MTMC datasets. Column "Opt. Flow" refers to the use of optical flow as additional inputs.

Method	**Opt Flow**	**Reference**	**MARS**		**Duke-MTMC**	
			mAP	Rank-1	mAP	Rank-1
LOMO+SQDA (Liao, Hu, Zhu & Li, 2015) Histogram based	No	CVPR 2015	16.4	30.7	-	-
Set2set (Liu, Junjie & Ouyang, 2017c) Custom Network	No	CVPR 2017	51.7	73.7	-	-
JST RNN (Zhou, Huang, Wang, Wang & Tan, 2017) CaffeNet	No	CVPR 2017	50.7	70.6	-	-
k-reciprocal (Zhong, Zheng, Cao & Li, 2017) ResNet50	No	CVPR-2017	58.0	67.8	-	-
TriNet (Hermans *et al.*, 2017a) ResNet50	No	ArXiv 2019	67.7	79.8	-	-
RQEN (Song *et al.*, 2018b) GoogLeNet	No	AAAI 2018	71.1	**77.8**	-	-
Part-Alignment (Suh, Wang, Tang, Mei & Lee, 2018b) GoogLeNet	No	ECCV 2018	72.2	83.0	78.3	83.6
Mask-Guided (Song *et al.*, 2018) ResNet50	No	CVPR 2018	71.1	77.1	-	-
Snippet (Chen, Li, Xiao, Yi & Wang, 2018a) ResNet50	Yes	CVPR 2018	71.1	82.1	-	-
STA (Fu, Wang, Wei & Huang, 2019) ResNet50	No	AAAI-2019	80.8	86.3	94.9	96.2
RevstTemPool (Gao & Nevatia, 2018b) ResNet50	No	Arxv 2018	75.8	83.1	-	-
Cosam (Subramaniam *et al.*, 2019b) ResNet50	No	ICCV 2019	76.9	83.6	93.5	93.7
STAL (Chen *et al.*, 2019) ResNet-50	No	IEEE TIP 2019	73.5	82.2	-	-
Cosam (Subramaniam *et al.*, 2019b) SE-ResNet50	No	ICCV 2019	79.9	84.9	94.1	95.4
STAR (Wu, Zhu & Gong, 2019b) ResNet-50	Yes	BMVC 2019	76.0	85.4	-	-
SCAN (Zhang *et al.*, 2019b) ResNet-50	No	IEEE TIP 2019	76.7	86.6	-	-
SCAN (Zhang *et al.*, 2019b) ResNet-50	Yes	IEEE TIP 2019	77.2	87.2	-	-
GLTR (Li, Wang, Tian, Gao & Zhang, 2019b) ResNet-50	Yes	CVPR 2019	78.5	87	93.7	96.3
M3D (Li, Zhang & Huang, 2020) ResNet-50	No	IEEE TIP 2020	79.46	88.63	93.67	95.49
Rec3D (Chen, Lu, Yang & Zhou, 2020a) ResNet-50	Yes	IEEE TIP 2020	80.4	86.3	-	-
MultiGrain (Zhang, Lan, Zeng & Chen, 2020c) ResNet-50	No	CVPR 2020	**85.9**	**88.8**	-	-
Mutual Attention ResNet-50	Yes	**Ours**	80.0	86.6	**94.9**	95.6
Mutual Attention SE-ResNet-50	Yes	**Ours**	80.9	87.3	94.8	**96.7**

Table 3.5 Accuracy of our proposed vs state-of-the-art methods evaluated on the ILIDS-Vid dataset. "Opt. Flow" refers to the use of optical flow as additional inputs.

Method	Opt. Flow	Reference	Rank 1	Rank 5
Two Stream* (Chung *et al.*, 2017)	Yes	ICCV 2017	60.0	86.0
Snippet (Chen *et al.*, 2018a) ResNet50	Yes	CVPR-2018	85.4	87.8
RQEN (Song *et al.*, 2018b) GoogLeNet	No	AAAI-2018	77.1	93.2
STAL (Chen *et al.*, 2019) ResNet-50	No	IEEE TIP	82.8	95.3
COSAM SE-ResNet50 (Subramaniam *et al.*, 2019b)	No	ICCV-2019	79.6	95.3
GLTR (Li *et al.*, 2019b) ResNet-50	Yes	CVPR-2019	86.0	-
Rec3D (Chen *et al.*, 2020a) ResNet-50	Yes	IEEE TIP-2020	87.9	98.6
Mutual Attention ResNet-50	Yes	Ours	86.2	96.4
Mutual Attention SE ResNet-50	Yes	Ours	**88.1**	98.4

Table 3.6 A comparison of attention methods and time complexity (in Flops) of some SOA methods

Model	Attention Method	Optical Flow	Time Complexity
Mask Guided (Song *et al.*, 2018)	Video based + binary segmented mask.	No	N/A
COSAM (Subramaniam *et al.*, 2019b)	Video co-segmentation based	No	35 G
GLTR (Li *et al.*, 2019b)	Video Based temporal self attn	Yes	60 G
SCAN (Zhang *et al.*, 2019b)	Video based inter-sequence attn	Yes	70 G
STAR (Wu *et al.*, 2019b)	Video frame level association based	Yes	N/A
REC3D (Chen *et al.*, 2020a)	3D attn. by Reinforcement Learning	Yes	N/A
Snippet (Chen *et al.*, 2018a)	Video LSTM based aggregation	Yes	60 G
MA (Ours)	Two stream mutual attention	Yes	58 G

close to the other SOA methods although COSAM (Subramaniam *et al.*, 2019b) is more efficient comparatively.

3.5.5 Visualization:

Fig. 3.5 shows activations of feature maps from final layer of our backbone 2D-CNN feature extractor on some bounding box images from the MARS dataset. The first columns shows original input image, the second column shows the corresponding optical flow, the third column shows activation after our proposed attention. The fourth column shows the activation map of our baseline that was generated by training our backbone without relying on optical flow, but that preserves the other settings. This was chosen as our baseline because the SOA methods use different backbones and different experimental settings. Our baseline is similar to (Gao & Nevatia, 2018b). It can be observed from the examples shown in Fig. 3.5 that our proposed flow guided mutual attention enhanced the spatial activations of feature maps based on the optical flow produced by the virtue of motion of persons between different consecutive frames of the sequence. Using the proposed approach, some additional regions of the person are activated (corresponding to moving parts), while the baseline approach only focuses mostly on a potentially discriminant spatial region.

It can be observed that the overall noise from the background has been suppressed particularly in the last row where the person is poorly localised. Fig. 3.6 shows a tracklet with its activations and the last row shows single tracklet feature 1) by proposed weighted addition and 2) average pooled. It can be observed that our proposed weighted addition method captures activation well from salient regions in the tracklet (Shows a good response for both head and legs compared to average pooled version).

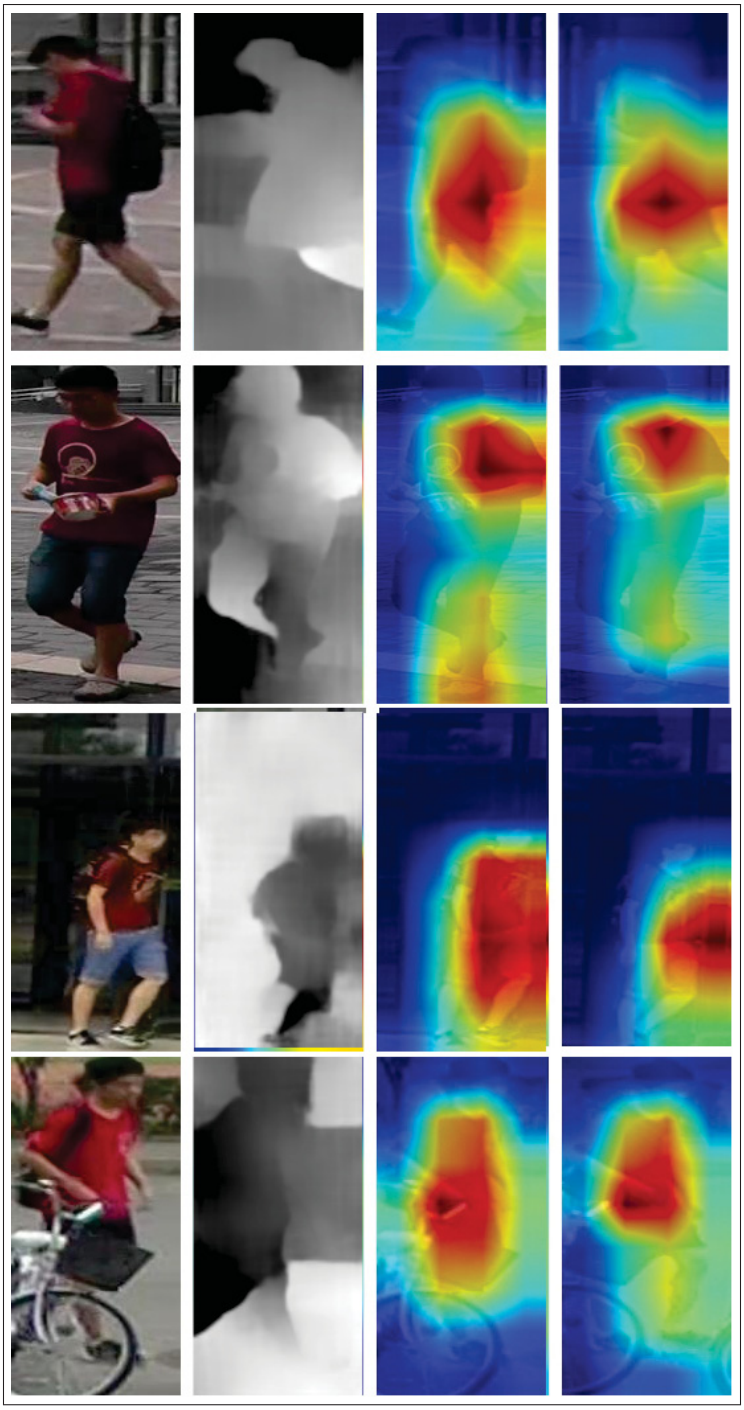

Figure 3.5 Visualization of feature maps produced using bounding box images from the MARS dataset. The first column shows the original input image, the second column shows corresponding optical flow, and the third column shows the corresponding activation maps of our proposed attention. The fourth column shows the activation map for our baseline, generated by training using our backbone without using the optical flow and keeping the rest of the setting same

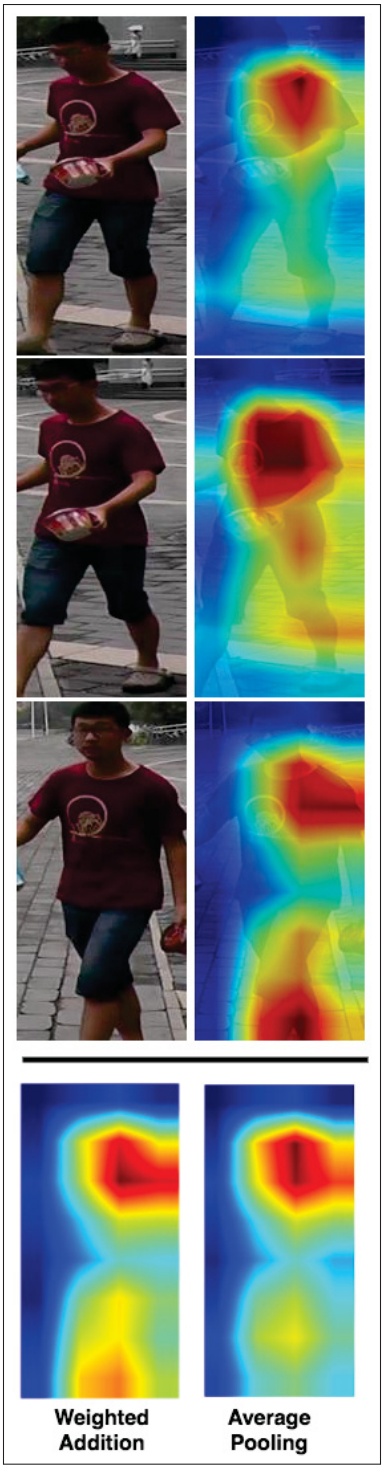

Figure 3.6 Visualization of features of a tracklet from the MARS dataset and our proposed weighted addition of the features vs average pooling of the features to extract one aggregated feature for the tracklet. – The last row shows our proposed weighted addition for the features of the tracklet, and the average pooled version of the tracklet

3.6 Conclusion

In this work we present a novel framework for flow guided attention and temporal feature aggregation for Person-ReID. We focus on visual appearances across spatial and temporal streams and their correlations to encode common saliency between the streams, reduce background clutter, learn motion patterns of person and to have the advantages of having longer sequences. Our feature aggregation method uses cues from both image and optical flow feature to assign weights and aggregate image instance features to produce a single video feature representation unlike assigning equal weights to images instances as in temporal pooling. Our method outperforms the state-of-the-art person reID methods in terms of both mAP and Rank1 accuracy evaluated on MARS, Duke-MTMC, and ILIDS-VID datasets. The proposed Mutual Attention network is most effective when the person detection and tracking produces high quality bounding-boxes, and in scenarios with bigger-sized bounding boxes for objects, where the attention helps in locating the objects.

CHAPTER 4

HOLISTIC GUIDANCE FOR OCCLUDED PERSON RE-IDENTIFICATION

Madhu Kiran[a] , Eric Granger[a] , R Gnana Praveen[a] , Le Thanh Nguyen-Meidine[a] , Soufiane Belharbi[a] , Louis-Antoine Blais-Morin[b]

[a]Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure,
1100 Notre-Dame St W, Montreal, Quebec H3C 1K3, Canada
[b]Genetec Inc., Montreal, Canada

Paper accepted for publication, Image and Vision Computing, October 2021

Abstract

In real-world video surveillance applications, person re-identification (ReID) suffers from the effects of occlusions and detection errors. Despite recent advances, occlusions continue to corrupt the features extracted by state-of-art CNN backbones and thereby deteriorate the accuracy of ReID systems. To address this issue, methods in the literature rely on an additional costly process, such as pose estimation, where pose maps provide supervision to focus on visible parts of occluded regions. In contrast, we introduce a Holistic Guidance (HG) method that relies on holistic (or non-occluded) data and its distribution in the dissimilarity space to train the CNN backbone on an occluded dataset. This method is motivated by our empirical study, where the distribution of pairwise between-class and within-class matching distances (Distribution of Class Distances or DCDs) between images has considerable overlap in occluded datasets compared to holistic datasets. Hence, our HG method employs this discrepancy in DCDs of both datasets for joint learning of a student-teacher model to produce an attention map that focuses primarily on visible regions of the occluded images. In particular, features extracted from both datasets are jointly learned using the student model to produce an attention map that allows dissociating visible regions from occluded ones. Additionally, a joint generative-discriminative CNN backbone is trained using a denoising autoencoder such that the system can self-recover from occlusions. Extensive experiments on several challenging public datasets indicate that the proposed approach can outperform state-of-the-art methods on both occluded and holistic datasets.

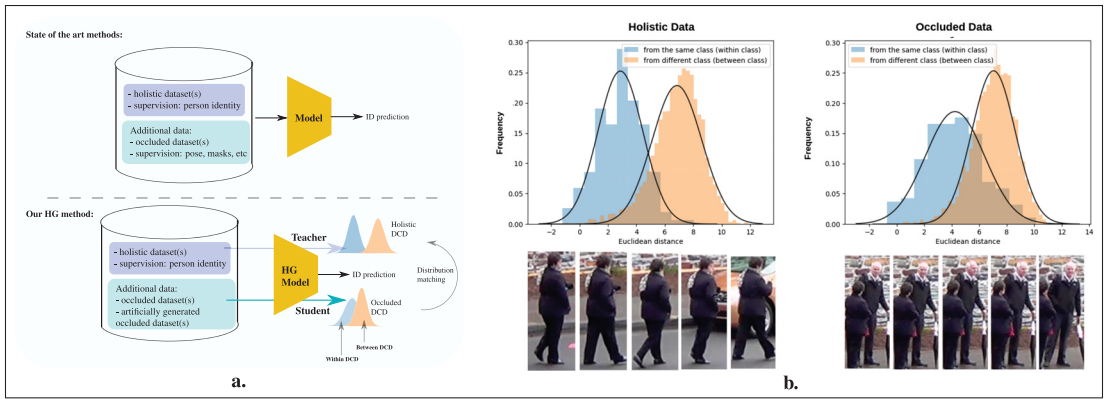

Figure 4.1 **(a)** An illustration of approaches to address occlusion in person ReID during training. **Top**: State-of-the-art models require additional supervision and occluded datasets. **Bottom**: Our proposed HG method requires no additional supervision but relies only on an additional holistic dataset for reference to non-corrupted features. **(b)** Examples of Class Distance Distributions of Duke-MTMC (**left**) and Occluded-Duke-MTMC (**right**) datasets measured in the distance space. The **blue DCD** shows the within-class distribution, while the **orange DCD** shows the between-class distribution. For Occluded-Duke-MTMC, within-class distances are relatively high and overlap with between-class distances

4.1 Introduction

Person Re-Identification (ReID) systems seek to associate individuals captured over a distributed set of non-overlapping camera viewpoints. This key visual recognition task has recently drawn significant attention due to its wide range of applications, e.g., autonomous driving, pedestrian tracking, sports analytics, and video surveillance (Farenzena *et al.*, 2010; Panda *et al.*, 2017; Ristani, Solera, Zou, Cucchiara & Tomasi, 2016a; Zhang, Xiang & Gong, 2016). Despite the recent progress with deep learning (DL), person ReID remains a challenging task in real-world applications due to the non-rigid structure of the human body, variability of capture conditions (e.g., illumination, scale, motion blur), in addition to person detection issues like miss-alignment, background clutter, and occlusion (Zhuo, Lai & Chen, 2019; He *et al.*, 2019; Miao *et al.*, 2019).

This paper focuses on the occlusion issue for person ReID, a challenge that has attracted much attention (Gao *et al.*, 2020a; He, Liang, Li & Sun, 2018a; He *et al.*, 2019; Miao *et al.*, 2019; Tan, Liu, Tian, Yin & Li, 2020; Wang *et al.*, 2020a). When bounding boxes or regions of interest (ROI) are occluded, the CNN backbone extracts noisy features, leading to pairwise matching

errors between query and reference ROIs and poor ReID accuracy for the occluded class. Since occlusions are diverse in color, shape, and size, extracting features from the entire ROI can potentially corrupt the global representation.

Several authors have attempted to address occluded person ReID by using pedestrian detectors that can additionally refine person ROIs (Zhang, Wen, Bian, Lei & Li, 2018a). Other methods follow an intuitive solution of masking occluded regions, extracting occlusion aware features, or applying weights and masks to occluded regions, applying body masks based on pose estimation, etc. (Gao *et al.*, 2020a; Miao *et al.*, 2019; Cai, Wang & Cheng, 2019a). With these state-of-the-art methods, the external mechanism for mask generation add a considerable time complexity during run-time. Unlike these methods, we propose a new method that only requires person identity labels as supervision and does not rely on additional supervision such as pose estimation. This provides robustness to occlusion in person ReID, yet lower complexity during inference.

Our method is motivated by the fact that the distribution of features learned from occluded and holistic[1] ReID datasets are different. The dissimilarity space can be regarded as space defined by dissimilarity coordinates, and CNN features are transformed into that space by computing pairwise matching distances for within-class and between-class samples in a given batch. It has been shown to successfully learn to separate feature representations for data that is noisy and overlapping (Jacobs, Weinshall & Gdalyahu, 2000; Duin, Bicego, Orozco-Alzate, Kim & Loog, 2014). Occluded person ReID is a good example of a problem with class overlap. Inspired by (Costa, Bertolini, Britto, Cavalcanti & Oliveira, 2020b; Elhamifar, Sapiro & Vidal, 2012; Ekladious, Lemoine, Granger, Kamali & Moudache, 2020; Jacobs *et al.*, 2000), we consider the dissimilarity space to capture the discrepancy between images in occluded and holistic datasets.

Fig. 4.1 shows the distribution of within-class and between-class distances (DCDs) among pairs of samples extracted from occluded and holistic datasets. We note two aspects: **(1) Within-class DCDs**: In an occluded dataset, this distance tends to be greater, with high variance, compared

[1] In the person ReID, the term "holistic" refers to image data that contains the full body of a person in both query and gallery sets. Holistic (or non-occluded) datasets have fewer occluded samples w.r.t. the overall dataset size.

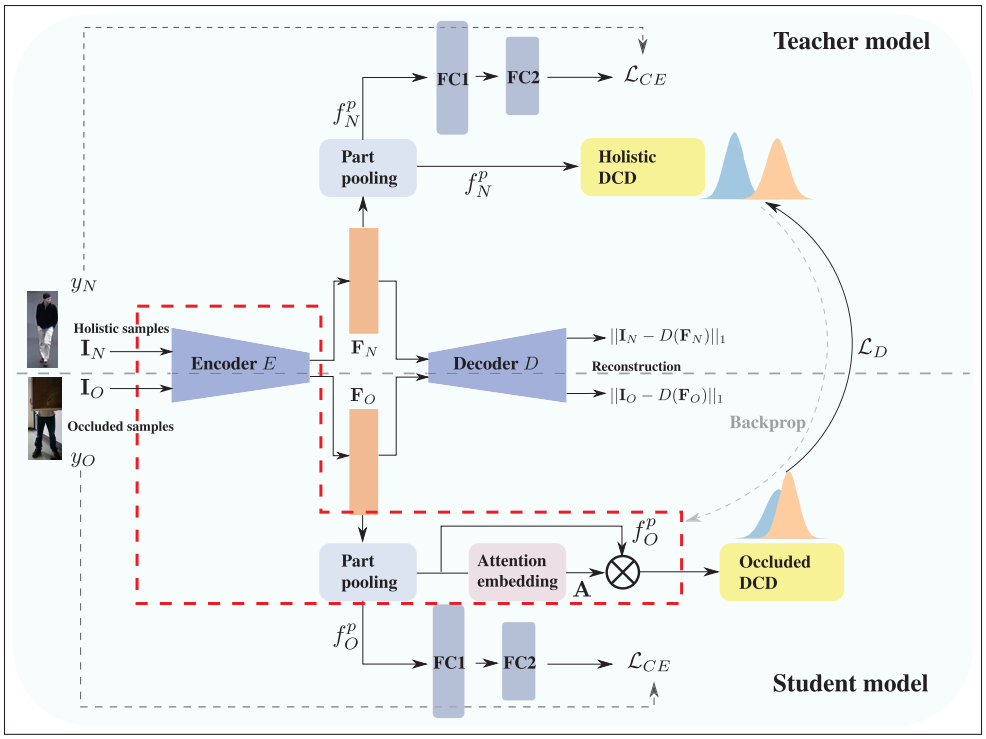

Figure 4.2 Our proposed HG method where a teacher model uses a holistic data distance distribution to train the student network(trained on artificially occluded or real occluded samples) such that it can accurately recognize persons appearing in occluded images

to holistic or non-occluded cases. Such distances are normally expected to be lower due to the similarity among samples within the same class. **(2) Overlap within- and between-class DCDs**: Large within-class DCDs are caused by samples being pushed away from the same class, allowing for substantial overlap with samples from other classes. The overlap between samples of different classes is more likely to impede discrimination among classes, leading to poor recognition. Note that both datasets used to generate Fig.4.1 – Duke-MTMC (Holistic) (Ristani *et al.*, 2016a) and Occluded-Duke (Miao *et al.*, 2019) – are from the same domain, but occlusions still cause such discrepancies. Based on these observations of DCDs, and considering the poor performance of models over occluded datasets, we hypothesize that occlusion is a potential source of corruption for feature representations in person ReID tasks.

This paper proposes a Holistic Guidance (HG) student-teacher network that relies on the distribution of holistic data in the dissimilarity space to train a student (CNN backbone) on an occluded dataset. The discrepancy of within- and between-class DCDs across datasets allows the network to extract features on occluded samples while simultaneously maintaining a good between-and within-class separation. Models trained on occluded data tend to overfit due to class overlap, so we advocate for using guidance from non-corrupted features of a larger holistic data in the dissimilarity space to mitigate this issue. Although both datasets can have different identities, transforming the samples to the dissimilarity space translates into a binary classification problem (Jacobs *et al.*, 2000; Costa *et al.*, 2020b). In practice, such learning scenarios could be achieved by using a single holistic dataset, and by building an artificially (augmented) occluded dataset. A second alternative involves using a training set consisting of real occluded samples. Our method performs HG since it relies on features with properties learned from holistic data to guide the CNN backbone in learning features of an occluded dataset.

Our method (shown in Fig.4.2) is also comprised of a shared generative model (i.e., a denoising auto-encoder) that is trained simultaneously on both datasets to enable self-recovery from occlusion. The student model has an additional embedding for producing an attention map, allowing the partial or local features to re-focus attention only on visible body-part features while ignoring the occluded regions that cause distribution discrepancy. Several authors have proposed generative models for person ReID (Wu *et al.*, 2019a; Zheng *et al.*, 2019c), mainly for GAN based data augmentation. In contrast, we introduce a denoising auto-encoder as the CNN backbone for our HG student-teacher network, allowing to self-recover in cases of occlusion.

Main contributions: (1) A novel HG student-teacher network that relies on the distribution of holistic data in the dissimilarity space to train a CNN backbone on the occluded dataset. (2) To motivate our HG method, we show that within-class DCDs of Occluded-ReID datasets overlap with between-class DCDs by a larger margin than holistic ReID datasets, even in cases where the occluded dataset is a subset of a holistic one. (3) Extensive experiments were performed on challenging Occluded (Zhuo, Chen, Lai & Wang, 2018a; Miao *et al.*, 2019), Partial (Zheng,

Gong & Xiang, 2011; Zheng *et al.*, 2015c), and Holistic (Zheng *et al.*, 2015a; Ristani *et al.*, 2016a) ReID datasets show that our HG method can outperform many SOTA methods.

4.2 Related Work in Person-ReID

Image-Based Methods: Siamese Networks have first been used in (Yi *et al.*, 2014) that employs three Siamese sub-networks for deep feature learning. Most of the further work based on deep-architecture ReID (Ahmed *et al.*, 2015; Cheng *et al.*, 2016; Chen *et al.*, 2017a; Liu *et al.*, 2017a; Varior *et al.*, 2016b) approaches introduce an end-to-end ReID framework, where both feature embedding and metric learning have been investigated. A few attention-based approaches for deep ReID (Li *et al.*, 2017a; Su *et al.*, 2017; Zhao *et al.*, 2017b) address misalignment challenges by incorporating a regional attention sub-network into a base re-ID model. With part-based methods, local features are extracted from different regions to enhance the discriminative power of the features. Suh et al. (Suh *et al.*, 2018a) extracted parts from the feature map and trained each part with separate classifiers. (Kalayeh, Basaran, Gökmen, Kamasak & Shah, 2018) used parts method to extract local features. In addition to this, other methods such as (Sun *et al.*, 2018; Zhao *et al.*, 2017b) have used part pooling with attention for a refined partial feature.

Occluded Person ReID: Occluded person ReID is different from person ReID or holistic person ReID because during test time, the probe images are often occluded as in real-world applications. Hence (Zhuo *et al.*, 2018a) have proposed to use a binary classifier to classify the images as occluded or not to distinguish occluded ones from holistic images. (Miao *et al.*, 2019) considered using pose guided feature alignment to align part features or local features. Similar to (Miao *et al.*, 2019), other works align local or part features by pose estimation like (Zheng, Huang, Lu & Yang, 2019a; Wei, Zhang, Yao, Gao & Tian, 2017). Similar to this, (Wang *et al.*, 2020a) use pose estimator to help predict key-points on body parts and use graph-based methods. The main disadvantage of pose-aligned methods is that they suffer from requiring additional supervision like the pose estimation step, which is error prone (Zhang *et al.*, 2017a). Recently, (Gao *et al.*, 2020a) use a method where, in the first step, they obtain discriminative features by using pose guided attention and then mine for visible parts. (He *et al.*, 2019) have used foreground-background mask instead of the pose. Recently, (Li *et al.*, 2021a) used only

identity information to learn occluded ReID with part aware transformers. Note that they add additional complexity with the transformers used for attention. (Jia *et al.*, 2021) is an other approach using only identity information, where Occluded ReID is considered as a set matching task to make invariant to occlusion of different parts.

Similar to our method, Zhuo et al. (Zhuo *et al.*, 2019) employ a student-teacher model. In the first stage, their teacher network learns from a holistic dataset, where both identification and supervised saliency is learned using a salient object detector. However, it differs from our approach in that: **(1)** it does not learn any supervised salient region detection, and **(2)** it allows recovering from small occlusions by using a denoising auto-encoder as a CNN backbone.

4.3 Proposed Approach

A Holistic Guidance (HG) method is introduced for unsupervised learning of attention maps in Occluded-ReID, without the need for any external guidance such as pose or segmentation maps. A student-teacher network is proposed, where the teacher network relies on holistic images or non-occluded images to teach the student network the DCD of holistic features. This allows the student to learn an attention map such that, when applied to occluded images, results in uncorrupted features like those of the teacher. More specifically, a joint generative and discriminative backbone is trained with a denoising autoencoder for the student-teacher model. It simultaneously learns to match image pairs while reconstructing images.

Problem Formulation: Let I_O and I_N denote input images from the occluded and holistic dataset, respectively. N and O denote the components from teacher and student trained from holistic data and occluded data, respectively. Let y_O and y_N denote the identity labels of occluded and holistic datasets. $\mathbf{F}_O$ and $\mathbf{F}_N$ represent the global feature maps of occluded and holistic datasets, which are obtained from the shared encoder E. The local or part-based features f_O^i, f_N^i are produced by applying a pooling function on $\mathbf{F}_O$ and $\mathbf{F}_N$. By minimising the discrepancy between-class distance distribution of f_O^i, f_N^i, we intend to learn an attention map A^i that is applied on the f_O^i to obtain features ${f_a}^i$ for $i = 1, .., p$ for each part. During testing, ψ, which is

a concatenation of global and local features, allows extracting features for matching from the gallery and query using a distance function, and to retrieve the identity.

Robust Backbone Model: We propose a joint learning framework of a denoising auto-encoder along with the classification networks to be robust to occlusions for person ReID. The input images are augmented by adding small noise using random erasing data augmentation, while the reconstruction loss is obtained using actual images. The encoder E is shared between denoising auto-encoder and ReID classification layers. In order to obtain robust features with both Generative and Discriminative properties, we reconstruct the input images using a Decoder on the embedding $\mathbf{F}_N$ and $\mathbf{F}_O$. Hence E and D together form a denoising auto-encoder. We have used a denoising autoencoder in order to exploit the full potential of the generative capability to take full advantage of the class distributions of the holistic data-set. Let $\widehat{\mathbf{I}_\mathbf{c}}$ represent the reconstructed image, and $\mathbf{F_c} = E(\mathbf{I_c})$ be the latent feature representation of the encoder, where $c \in \{N, O\}$ (holistic and occluded images) and $\mathbf{I_c}$ is the input image. The size of $\mathbf{F_c}$ is $B \times C \times w \times h$, where B is the batch, C is the number of output channels of the encoder E and w, h width and height of the feature map. $\mathbf{F_c}$ can also be referred to as the latent feature representation of the auto-encoder. A part-based pooling is applied on $\mathbf{F_c}$ to obtain p parts of stripes of features. Our part pooling method used is similar to (Sun *et al.*, 2018) where a 2D feature map is split into horizontal stripes of p parts. Global average pooling is then performed on each of the p parts to obtain p feature vectors of size C. Each feature vector of p parts is assigned to a unique classifier, resulting in p classifiers trained using identity labels of the corresponding datasets.

The predicted output for each given image $\mathbf{I_c}$ from the classifier is $\hat{y}_{i,c}$, where $i = 1, \ldots, p$ parts. The identity prediction loss function over all the parts is:

$$\mathcal{L}_{\text{CE,c}} = \frac{1}{K} \sum_{i}^{K} -\log \left(\frac{\exp\left(\mathbf{W}_{y_i}^T \mathbf{x}_i + b_{y_i}\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{W}_j^T \mathbf{x}_i + b_j\right)} \right) \tag{4.1}$$

where $\mathcal{L}_{\text{CE,c}}$ is the cross entropy loss, K is the batch size, class label $y_i \in \{1, 2, \ldots, N\}$ is associated with i^{th} training image. $\mathbf{W}_j$ and b_j are the weights and bias of last fully connected layer for class y. Similarly $\mathbf{W}_j$ and b_j are the weights and bias of the the j^{th} class. The denoising auto-encoder is learned using reconstruction loss between the reconstructed and original images before random erasing augmentation (Zhong, Zheng, Kang, Li & Yang, 2020b), which acts like noise added in denoising auto-encoders. We propose this inspired by (Cheng, Wang, Gong & Hou, 2015) to self recover from occlusion. Normally in (Cheng *et al.*, 2015) an autoencoder is learnt to generate non occluded face image from artificially occluded face image which is then post processed for recognition. But in our case since we use a joint representation we expect the joint feature to have the self recovering properties. The reconstruction loss of the autoencoder is given by,

$$\mathcal{L}_{\text{recon,c}} = \mathbb{E}\left[\|\mathbf{I_c} - D\left(\mathbf{F_c}\right)\|_1\right], \tag{4.2}$$

where D denotes the decoder function of the denoising auto-encoder. We do not use the reconstructed image for any other steps. However, the reconstruction loss is optimised in order to enable the deep features to have generative properties and to self-recover from occlusion. The total loss for the joint learning of generative discriminative learning is:

$$\mathcal{L}_{\text{joint},c} = \mathcal{L}_{\text{CE,c}} + \lambda \mathcal{L}_{\text{recon},c}\ , \tag{4.3}$$

where λ is the trade-off parameter. **Student-Teacher Model:** A student-teacher model with the proposed backbone is shared between the teacher and student. In addition to the backbone, the student model carries an embedding to produce attention maps. From Fig. 4.1, it can be observed that a DCD obtained by comparing extracted deep features of occluded in-class images overlap with those of out-of-class distances by a large margin in comparison with the holistic dataset. The overlap of the DCD indicates corrupted features as in-class distance distribution must have good separation from that of out-of-class distribution. We further justify the use of holistic guidance by the following. One could learn a separation between classes in the feature space using a triplet or contrastive loss on an occluded dataset alone. Yet, the model may overfit on the occluded dataset due to class overlap. However, holistic datasets are much larger than

occluded ones. Therefore, they can provide a good generalization of non-corrupted DCD given that class overlap can be well dealt within the dissimilarity space. By matching the DCD, the student network can learn the class overlap of the teacher network (which is well separated).

Fig. 4.2 shows our overall architecture along with the backbone auto-encoder-based deep feature extractor. We simultaneously take two input images—one from the holistic dataset and the other from the occluded dataset. The extracted deep features are simultaneously optimized for identity loss by learning a set of two fully connected layers for classification. We use two separate classifiers, one for the teacher to learn holistic data identities and the other for the student to learn occluded data identities.

Attention Embedding is particular to student network alone capable of producing attention for the partial features such that the attended partial feature will have good separation between within and between-class distances similar to that of the teacher. Let the attention embedding (which is a set of two layers 1×1 convolutional filters with ReLU between them followed by batch normalization layer with sigmoid activation for final attention output) be represented by AE. The attention produced by the attention embedding is given by $A^i = AE(f_O^i)$. The attention maps in A^i of size $B \times C$ are obtained for each partial feature, with $i = 1, \ldots, p$ and p being the number of parts. The attention obtained is multiplied with each partial feature to obtain attended partial features, ${f_a}^i = f_O^i \bigotimes A^i$. The layers for occluded image classification, FC_c^i, are applied on each of the attended partial feature ${f_a}^i$. While training on an artificially occluded dataset alone, we use a binary classifier to learn occluded or non-occluded images on artificially occluded samples similar to (Zhuo *et al.*, 2018a).

In order to *learn the attention*, the student network relies on occluded input images and distance distribution matching. Then, DCD of the occluded and holistic features are compared. Given a mini-batch of image input with occluded and holistic images I_O and I_N, partial features f_N^i and ${f_a}^i$ (partial feature with attention) are extracted. We denote the class identity for the features by u and v. Therefore, for each mini-batch, we extract pairs of image features with different

combinations within a batch according to:

$$d_i^{\text{wc}}\left(\mathbf{I_c^u}, \mathbf{I_c^v}\right) = \left\|P_N^{i,u} - P_N^{i,v}\right\|_2, u = v \ , \ \text{ and, } \ d_i^{\text{bc}}\left(\mathbf{I_c^u}, \mathbf{I_c^v}\right) = \left\|P_N^{i,u} - P_N^{i,v}\right\|_2, u \neq v \ . \tag{4.4}$$

Eqn. 4.4 transforms the features to *dissimilarity space*. P^i denotes the part features , f_N^i for holistic data and $f_a{}^i$ for occluded data. The distance distributions are extracted from d_i^{wc} and d_i^{bc} for both holistic and occluded data. We implicitly learn to produce a good attention embedding AE by minimising the discrepancy between DCD of holistic and occluded data using MMD (Gretton, Borgwardt, Rasch, Schölkopf & Smola, 2012). Let $\mathbf{D_c^{wc}}$ and $\mathbf{D_c^{bc}}$ be the distributions from d_i^{wc} and d_i^{bc}. The losses measuring the discrepancy between class distributions of holistic and occluded data are,

$$\mathcal{L}_D^{\text{wc}} = MMD\left(\mathbf{D}_N^{\text{wc}}, \mathbf{D}_O^{\text{wc}}\right) \ , \ \mathcal{L}_D^{\text{bc}} = MMD\left(\mathbf{D}_N^{\text{bc}}, \mathbf{D}_O^{\text{bc}}\right) \ \text{ and, } \ \mathcal{L}_{\text{global}} = MMD\left(f_N, f_a\right) \tag{4.5}$$

It is important to note that losses $\mathcal{L}_D^{\text{wc}}$ and $\mathcal{L}_D^{\text{bc}}$ are optimized by fixing $\mathbf{D_N^{bc}}$ and $\mathbf{D_N^{wc}}$, the DCD of teacher network. This allows the student network distance distribution to match that of the teacher network. Hence, Eqn. 4.5 calculates the discrepancies between the within-class and between-class DCDs of holistic datasets and occluded datasets. This loss is minimized during learning to obtain a good attention map from the embedding to focus on non-occluded regions of occluded images. $\mathcal{L}_{global}$ calculates the MMD distance between teacher features and student features to encourage the model to perform well on both occluded and holistic data. This $\mathcal{L}_{global}$ is particularly used when the holistic and occluded datasets are from different domains. Parameters λ_1, λ_2, and λ_3 balance these losses, and are determined empirically.

$$\mathcal{L}_D = \lambda_1 \mathcal{L}_D^{\text{wc}} + \lambda_2 \mathcal{L}_D^{\text{bc}} + \lambda_3 \mathcal{L}_{\text{global}} \ . \tag{4.6}$$

End-to-End Learning and Testing: The full system is optimized for reconstruction losses and identity losses with cross-entropy on both occluded and holistic data, and finally, the class

distance distribution loss. The overall loss function $\mathcal{L}_{Total}$ is given by,

$$\mathcal{L}_{\text{Total}} = \alpha \mathcal{L}_{\text{joint},c} + (1 - \alpha)\mathcal{L}_{D} \ , \tag{4.7}$$

where α balances these losses, and is determined empirically (see supp. material). During testing, only the student is used along with the components denoted within Fig. 4.2 to extract ψ, which is a concatenation of global and local features, F_O and ${f_a}^i$, for both gallery and query images. Extracted features are matched with Euclidean or Cosine distance to retrieve the identity of a query image from the gallery.

4.4 Experimental Results and Discussion

Datasets: Our approach is validated on three challenging groups of datasets – Holistic ReID, Occluded-ReID, and Partial ReID datasets. Our main objective is to address performance on Occluded-ReID problems (Occluded-DukeMTMC (Miao *et al.*, 2019) and Occluded-ReID (Zhuo *et al.*, 2018a) datasets) and Partial ReID problems (Partial-ILIDS (Zheng *et al.*, 2011; Miao *et al.*, 2019) and Partial-ReID (Zheng *et al.*, 2015c) datasets), but we also evaluated on Holistic problems – Market1501 (Zheng *et al.*, 2015a) and Duke-MTMC (Ristani *et al.*, 2016a) datasets – to further assess the effectiveness our approach on regular ReID problems. The Occluded-DukeMTMC dataset (Miao *et al.*, 2019) contains a total of 15,618 training and 17,661 gallery with 2,210 occluded query images. This is a subset of the Duke-MTMC dataset. To test on Occluded-Duke, we train the student model with Occluded-Duke training set similar to (Miao *et al.*, 2019; Zhou, Wu, Zhang & Sehdev, 2020b; Wang *et al.*, 2020a). The Occluded-ReID dataset (Zhuo *et al.*, 2018a) mimic real-world application scenarios by collecting datasets using mobile camera equipment on campus. It has a total of 2,000 annotated images with 200 identities. Each identity consists of 5 full-body images and 5 partial images. The Holistic ReID and Partial ReID datasets are described in the supplementary material.

Implementation Details: For validation, the ResNet50 (He *et al.*, 2016) was implemented as our backbone Encoder. Transposed convolution layers were used along with Interpolation for

the Decoder (see details in the supplementary material). To evaluate Occluded-Duke-MTMC, we train the student using the train data of Occluded-Duke-MTMC, and the teacher with Market1501 (Zheng *et al.*, 2015a), as in the SOA. Since Occluded-ReID and Partial ReID do not have a prescribed set of training images, the whole dataset was used for testing (as in (Gao *et al.*, 2020a)). To have a common setting with the SOA (Miao *et al.*, 2019), we use an input image size of 384 X 128. We train our backbone with Partial Features with $p = 6$ parts, and set the co-efficient $\lambda = 0.01$. The teacher network is pre-trained for 15 epochs, and the student-teacher is trained together for 120 epochs. The Adam optimization is used with an initial learning rate of 0.0003. We report the Cumulative Matching Characteristics (CMC) and mean average precision (mAP) (Zheng *et al.*, 2015a).

Results with Occluded and Partial ReID Problems. Tabs. 4.1 and 4.2 show the result of our method on the Occluded-Duke and Occluded-ReID dataset compared with State-Of-The art (SOA) methods. (Miao *et al.*, 2019; Zhou *et al.*, 2020b; Wang *et al.*, 2020a) are Occluded-ReID methods. We show the results for our method with both ResNet50 and ResNet-IBN (Pan, Luo, Shi & Tang, 2018) backbones. In the Table "PM"-pose maps, "KP"-key -point detection. Our method outperforms all the other Occluded Person ReID methods mentioned in the table. Our proposed HG method performs competitively in Rank-1 without any external input, such as pose or segmentation masks, on the Occluded-Duke dataset and 2.5% on the Occluded-ReID dataset. Since the Occluded-ReID dataset does not contain training images, we show two sets of results,

Table 4.1 Accuracy of HG and state-of-the-art methods on the Occluded-Duke dataset

Method	**Backbone**	**Supervision**	**Accuracy**			
			Rank-1	Rank-5	Rank-10	mAP
LOMO+XQDA (Liao *et al.*, 2015), CVPR 2015	-	None	8.1	17.0	22.0	5.0
Part Aligned (Zhao *et al.*, 2017b), ICCV 2017	GoogLeNet	None	28.8	44.6	51.0	20.2
Random Erasing (Zhong *et al.*, 2020b), AAAI 2020	ResNet50	None	40.5	59.6	66.8	30.0
HACNN (Li, Zhu & Gong, 2018a), CVPR 2018	Custom	None	34.4	51.9	59.4	26.0
Adver Occluded (Huang, Li, Zhang, Chen & Huang, 2018a), CVPR 2018	ResNet50	None	44.5	-	-	32.3
PCB (Sun *et al.*, 2018), ECCV 2018	ResNet50	None	42.6	57.1	62.9	33.7
Part Bilinear (Suh *et al.*, 2018a), ECCV 2018	GoogLeNet	Occluded-Duke	36.9	-		-
PGFA (Miao *et al.*, 2019),ICCV, 2019	ResNet50	Occluded-Duke + PM	51.4	68.6	74.9	37.3
Depth Occln (Zhou *et al.*, 2020b), PRL 2020	ResNet50	Occluded-Duke + PM	53.0	67.0	72.9	38.1
HOReID (Wang *et al.*, 2020a), CVPR 2020	ResNet50	Occluded-Duke + KP	55.1	-	-	43.8
PAT (Li *et al.*, 2021a), CVPR 2021	ResNet50	Occluded-Duke	64.5	-	-	53.6
MOS (Jia *et al.*, 2021), AAAI 2021	ResNet50	Occluded-Duke	61.0	-	-	49.2
MOS (Jia *et al.*, 2021), AAAI 2021	ResNet50-IBN	Occluded-Duke	**66.6**	-	-	**55.1**
HG(ours)	ResNet50	Occluded-Duke	61.4	77.0	79.8	50.5
HG(ours)	ResNet50-IBN	Occluded-Duke	65.1	**79.1**	**81.4**	54.7

Table 4.2 Accuracy of HG and state-of-the-art methods on the Occluded-ReID dataset

Method	Backbone	**Supervision**	**Accuracy**			
			Rank-1	Rank-5	Rank-10	mAP
IDE (Zheng *et al.*, 2015a), ICCV 2015	-	None	52.6	68.7	76.6	46.4
MLFN (Chang, Hospedales & Xiang, 2018), CVPR 2018	Custom	None	42.3	60.6	68.5	36.0
HACNN (Li *et al.*, 2018a),CVPR 2018	Custom	None	29.1	44.7	54.7	26.1
PCB (Sun *et al.*, 2018), ECCV 2018	ResNet50	None	59.3	75.2	83.2	53.2
Part Bilinear (Suh *et al.*, 2018a), ECCV 2018	GoogLeNet	None	54.9	70.8	77.7	50.3
teacher-S (Zhuo *et al.*, 2019), ArXiv 2019	ResNet	Split Test Set	55.0	64.5	77.3	59.8
PGFA (Miao *et al.*, 2019), ICCV 2019	ResNet50	PM	57.1	77.9	84.0	56.2
PVPM (Gao *et al.*, 2020a), CVPR 2020	ResNet50	PM	70.4	84.1	89.8	61.2
HOReID (Wang *et al.*, 2020a), CVPR 2020	ResNet50	KP	80.3	-	-	70.2
PAT (Li *et al.*, 2021a)	ResNet50	-	81.6	-	-	72.1
HG (ours Unsup)	ResNet50	None	79.4	88.5	93.7	71.1
HG (ours Sup)	ResNet50	Occluded-Duke	82.3	89.7	94.1	71.7
HG (ours Sup)	ResNet50-IBN	Occluded-Duke	**82.8**	**90.1**	**94.6**	72.0

one with student trained on artificially occluded Market1501 dataset -HG(Unsup) and the other the student model trained on occluded samples from Occluded-Duke dataset -HG(Sup). Our results show that student trained on augmented Market1501 already outperforms many SOTA. Additionally, when the student is trained on a totally different occluded dataset (Occluded-Duke) from that of the test dataset (Occluded-ReID) still outperforms all SOTA. We also evaluate our method on partial ReID datasets (Zheng *et al.*, 2015c) and the result table is presented in the supplementary material. We can observe from the results that our method has improved over the SOTA by 2%.

Table 4.3 Impact on HG accuracy of training data and architecture on Occluded-ReID dataset

Model	**Experiment**	**Training Data (Student)**	**Rank-1**
Impact of Training Architecture (Test Set: Occluded-ReID)			
HG (student+teacher)	w/o occlusion classifier	Augment Holistic	75.6
	w/ occlusion classifier	Augment Holistic	79.4
	w/o occlusion classifier	Occluded-Duke	**82.3**
Impact of Losses (Test Set: Occluded-ReID)			
teacher only	identity only	-	59.5
teacher+student	identity only	Occluded-Duke	63.8
teacher+student	identity+reconstruction	Occluded-Duke	67.7
teacher+student	identity+reconstruction+MMD	Occluded-Duke	76.1
teacher+student	MMD + attention (no autoencoder)	Occluded-Duke	80.1
teacher+student	reconstruction+MMD+attention	Occluded-Duke	**82.3**
Impact of backbone (Only ResNet50 Backbone, w/o Part Pooling)			
teacher only	identity	-	51.3
teacher+student	reconstruction+MMD+attention	Occluded-Duke	68.5

Table 4.4 Accuracy of our HG method on Market1501 and Duke-MTMC datasets

Model	**Backbone**	**Market1501**		**Duke-MTMC**	
		Rank-1	mAP	Rank-1	mAP
PCB (Sun *et al.*, 2018), ECCV 2018	ResNet50	92.3	77.4	81.8	66.1
MaskReID (Song *et al.*, 2018a), CVPR 2018	MSCAN	90.0	75.3	-	-
FPR (He *et al.*, 2019), ICCV 2019	ResNet50	95.4	**86.6**	88.6	**78.4**
PVPM (Gao *et al.*, 2020a), CVPR 2020	ResNet50	93.0	80.8	83.6	72.6
PGFA (Miao *et al.*, 2019), ICCV 2019	ResNet50	91.2	76.8	82.6	65.5
HOReID (Wang *et al.*, 2020a), CVPR 2020	ResNet50	94.2	84.9	86.9	75.6
MOS (Jia *et al.*, 2021), CVPR 2021	ResNet50-IBN	95.4	89.0	**90.6**	**80.2**
PAT (Li *et al.*, 2021a), CVPR 2021	ResNet50	95.4	88.0	88.8	78.2
HG (Ours)	ResNet50	**95.6**	86.1	87.1	77.5

Ablation Study. An extensive study was conducted on Occluded-ReID to analyze the impact on training data and architecture performance. Results are shown in Tab. 4.3.

(a) Impact of Data. This study was performed on the full student-teacher model with all the components. In Tab. 4.3, "Augm. on Holistic": refers to the student model trained with artificially occluded Market1501 dataset by Random Erasing (Zhong *et al.*, 2020b). We also use additional "occluded or non-occluded" binary classification on the features and show results with and without this classification. We can see that augmentation has helped the student model to learn Occluded-ReID and perform better than many SOTA. Using the Occluded-Duke to train the student shows that the student model performs even better on the Occluded-ReID dataset. Also, Occluded-Duke and Occluded-ReID datasets have no overlap.

(b) Impact of Architecture. To analyze the effect of architecture on the result from above, the teacher alone (pre-trained on Market1501) is fine-tuned on Occluded-Duke. Results indicate that fine-tuning the baseline network on Occluded-Duke performs poorly on Occluded-ReID. In Tab 3, the remarks "identity+reconstruction" refer to the experiment where the student-teacher model has been trained with identity loss and reconstruction loss alone. The reconstruction loss has helped the student to learn some generative properties, and hence the overall result on Occluded-ReID data is better than the baseline fine-tuned on the Occluded-Duke dataset. We further extend our experiments by using distribution loss, identity loss along with reconstruction loss, but no attention "reconstruction+MMD." The "MMD+attention" experiment does not use reconstruction loss, and hence the autoencoder is not trained. Finally, we train the student model

with the full system "reconstruction+MMD+attention," which includes the attention mechanism learned by matching distributions.

(c) Impact of CNN Backbone. In order to assess the impact of part pooling, we perform another the "Impact of backbone" experiment, where only a ResNet-50 backbone is used with no part pooling. Although the overall results are this case is lower than when using part pooled features, it can be seen that proposed loss has still improves Occluded-ReID performance. From the results, we can conclude that our system learns occluded-ReID by using either artificially-occluded examples or real-occluded examples with holistic data as a reference. We show additional ablation studies in the supplementary materials.

(d) Holistic ReID Datasets. Tab. 4.4 shows results with the HG model on holistic Market1501 (Zheng *et al.*, 2015a) and Duke-MTMC-(Ristani *et al.*, 2016a) datasets. Results show that HG is competitive with other SOTA models designed to work with the Occluded-ReID problem.

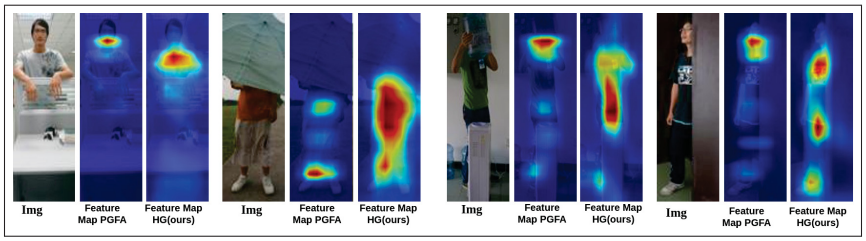

Figure 4.3 Activation maps generated for four occluded images of the Partial ReID dataset by the PGFA method (Miao *et al.*, 2019) versus our HG method. PGFA uses pose estimation additionally

Qualitative Results. Fig. 4.3 compares activation maps from the student model tested on examples with Partial-ReID dataset with activation maps of (Miao *et al.*, 2019) (uses pose maps for attention). In the figure, both head and legs are occluded. But, from the activation maps of our proposed HG method, it can be seen that our method is good at localising non-occluded regions alone.

4.5 Conclusion

This paper proposes a novel HG student-teacher model for occluded person ReID that only requires image identity labels but no costly process to focus on visible parts of occluded regions. The proposed HG teacher considers the DCD among samples in a holistic dataset to train a student model to generate attention maps, thereby alleviating the occlusion problem. Unlike most methods in the literature that use external supervision (such as pose) to create visibility cues, we only rely upon the distribution of holistic data during training, using it as a soft label. Hence, during test time, our model requires no external cues such as pose, and the overall parameters include just the backbone Encoder and a small embedding to generate attention maps to be used during feature extraction. Joint learning of a denoising autoencoder was used to improve the ability to self-recover from occlusion. Results on several challenging datasets show that our HG method can outperform state-of-the-art models for Occluded-ReID and Holistic ReID tasks.

CHAPTER 5

PERSON RE-IDENTIFICATION IN THE DISSIMILARITY SPACE FOR REAL-TIME APPLICATIONS

Madhu Kiran[a] , Eric Granger[a] , Kartikey Vishnu[b] , Le Thanh Nguyen-Meidine[a] , Rafael Menelau Oliveira e Cruz[a]

[a]Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure,
1100 Notre-Dame St W, Montreal, Quebec H3C 1K3, Canada
[b]Delhi Technological University, New Delhi, India

Paper to be submitted to Pattern Recognition

Abstract

Although deep learning (DL) models have been successfully applied in person re-identification (ReID), their matching accuracy can significantly decline in real-time video surveillance applications due to the scarcity of training data and the compromises needed to deploy lightweight architectures. Given their reduced capacity, fine-tuning lightweight backbones on limited annotated data can lead to greater overlap and dispersion of class distributions in the feature space. Furthermore, standard ReID methods adopt Euclidean feature space mapping, where a distance metric with a margin ensures that similar classes are closer and dissimilar classes are farther. The Euclidean distance space can face challenges in high-dimensional manifolds, exacerbated by the problem of class imbalance and overlap. Finally, standard person ReID training with contrastive losses requires hard sample mining and heuristic hard sample selection, which may suffer from sample selection bias and sensitivity to hyperparameters. To address these challenges, a new end-to-end training approach called Dissimilarity ReID (DisReID) is proposed that leverages the dissimilarity space for accurate real-time person ReID. Specifically, we introduce dichotomy transformation to project features from the Euclidean to dissimilarity space. The system is trained to optimize the backbone and max-margin classifier jointly for linear separation of within- and between-class samples. Our experimental results[1] show the benefits of our proposed DisReID approach, particularly for real-time applications that require

[1] Code is available: https://anonymous.4open.science/r/DisReID-5224/

lightweight backbones fine-tuned with limited data. Our approach achieves state-of-the-art performance on challenging ReID benchmark datasets.

5.1 Introduction

Person Re-Identification (ReID) is a visual recognition task that has gained significant attention due to its wide range of applications in monitoring and video surveillance (Ye *et al.*, 2021), including multi-camera target tracking, pedestrian tracking in autonomous driving, access control in biometrics, and human-computer interaction communities (Sun *et al.*, 2018; Quan *et al.*, 2019; Tay *et al.*, 2019; He *et al.*, 2021). This image retrieval task involves matching images of individuals captured over a distributed set of non-overlapping camera viewpoints. Despite recent progress with deep learning (DL) models, person ReID remains a challenging task in real-world applications due to the nonrigid structure of the human body, the variability of capture conditions (e.g., pose, illumination, scale, resolution, and motion blur), occlusions, and background clutter (Hermans *et al.*, 2017a).

The challenges mentioned above are typically addressed using a backbone model (e.g., CNN of vision transformer) trained, using some deep metric learning (DML) loss to provide discriminative feature embeddings for input images, and a distance or correlation function for pair-wise matching between query and gallery images. This DML paradigm translates into pairwise losses that encourage small distances for pairs of samples from the same class and large distances for samples from different classes (Hermans *et al.*, 2017a; Mekhazni *et al.*, 2020; Zhou, Yang, Cavallaro & Xiang, 2019). It is common to use triplet loss or a combination of losses for learning a person's identity using cross-entropy and triplet loss (Hermans *et al.*, 2017a; He *et al.*, 2021). State-of-the-art approaches for image-based ReID (Ahmed *et al.*, 2015; Bhuiyan *et al.*, 2020, 2014, 2018; Farenzena *et al.*, 2010; Panda *et al.*, 2017; Sun *et al.*, 2018; Quan *et al.*, 2019; Tay *et al.*, 2019; He *et al.*, 2021) seek to match still images of individuals captured across a network of non-overlapping video cameras.

Typically, these approaches train a deep CNN using one or more of these surrogate losses. However, both of these losses have their problems. Classification loss requires a growing number of learnable parameters as the number of identities increases, most of which are discarded after training. Metric learning approaches with distance-based loss, such as triplet loss, require extensive hard sample mining for the loss function to be effective. Hard sample mining often suffers from additional issues, such as selection bias during training; label noises present in the hard samples can get amplified due to the selection bias and the lack of diversity when selecting hard samples (Xuan *et al.*, 2020; Hermans, Beyer & Leibe, 2017b). Moreover, Euclidean and cosine distances are often used in the literature to match images in the feature space based on the proximity of a query sample to labeled samples in the neighborhood.

Nevertheless, from (Weller-Fahy, Borghetti & Sodemann, 2014; Xing, Jordan, Russell & Ng, 2002), it can be observed that these distances may suffer from the non-linearity of the manifolds. The low inter-class variations of ReID and the problem with high dimensional feature space for classification can cause overlaps. These problems can be observed in Fig 5.1. Fig. 5.1(a) shows the distribution of pair-wise distances of deep embeddings extracted from within- and between-class samples. The figure shows the distribution before (left column) and after (right column) training. From Fig. 5.1(a), it can be observed there is some overlap between the distribution of within-class and between-class distances, indicating the region of possible misclassification due to training in the Euclidean space.

At the same time, some of these problems can be overcome with very deep CNNs with an increasing number of parameters and better discriminative representation (Szegedy, Ioffe, Vanhoucke & Alemi, 2017). Recently, many authors have explored vision transformers (ViTs) (He *et al.*, 2021; Lai, Chai & Wei, 2021; Yu *et al.*, 2022a) and Self-Supervised pre-training for Person-ReID (Ye, Hong, Zeng & Zhuang, 2022; Ma, Li, Yuan & Zhao, 2023), and improving the state-of-art by a large margin. However, these methods suffer from high computational complexity and may not be suitable for real-time performance. Efficient ReID methods can significantly reduce the computational complexity and processing time required to identify individuals in a large-scale surveillance system. OSNet (Zhou *et al.*, 2019; Remigereau *et al.*,

2022b) has proposed omni-scale features for efficient person ReID, providing a good trade-off between accuracy and complexity. Other methods include (Li, Shao, Niu & Xue, 2021b; Yan *et al.*, 2021; Remigereau *et al.*, 2022a) propose strategies to improve the complexity-accuracy trade-off by architecture changes, knowledge distillation methods, etc. But these methods propose a completely new backbone as in OSNet (Zhou *et al.*, 2019) instead of a generic solution that can be used with the backbone of choice. Methods proposed for a specific backbone limit the choices of end applications.

This paper proposes a ReID framework, DisReID, to jointly learn a max-margin-classifier and the backbone feature extractor in the dissimilarity space for real-time applications. We propose to improve the intra-class variance of embeddings by transforming them from the embedding space to the dissimilarity space and dichotomy transformation. We classify pair-wise examples obtained from dichotomy transformation with the max-margin classifier to classify a given set of query and gallery examples as within-class and between-class examples to find the right match.

The dissimilarity approach to pattern recognition has been previously discussed by (Duin, Loog, Pkalska & Tax, 2010; Oliveira; Costa, Bertolini, Britto, Cavalcanti & Oliveira, 2020a) wherein a multi-class problem is transformed into a binary one, reducing the difficulties caused by the lack of data samples and complexity arising from hard sample mining. Transforming the embeddings in the dissimilarity space projects within-class samples closer to the origin of the dissimilarity space and between-class samples farther away from the origin. Additionally, (Costa *et al.*, 2020a) also points out that overlapping non-linear multi-class data in the feature space can be transformed into linearly separable data in the dissimilarity space. Thus simplifying the problem. Most of the literature related to dissimilarity space transform works with classical features (Duin *et al.*, 2010; Oliveira; Costa *et al.*, 2020a). At the same time, a few works use deep CNN embeddings in the dissimilarity space (Souza, Oliveira, Cruz & Sabourin, 2021), which use pre-trained Deep CNN features and transform them into dissimilarity space. They further perform feature selection using Binary particle Swarm optimization for feature selection to remove redundant features. However, they do not train end-to-end.

We hypothesize that when the classifier is trained end to end along with the backbone feature extractor, the samples can be optimally separated with reduced overlap. Our hypothesis has been inspired by the observation in Fig 5.1. Fig 5.1 (b) shows the classification of pairwise samples

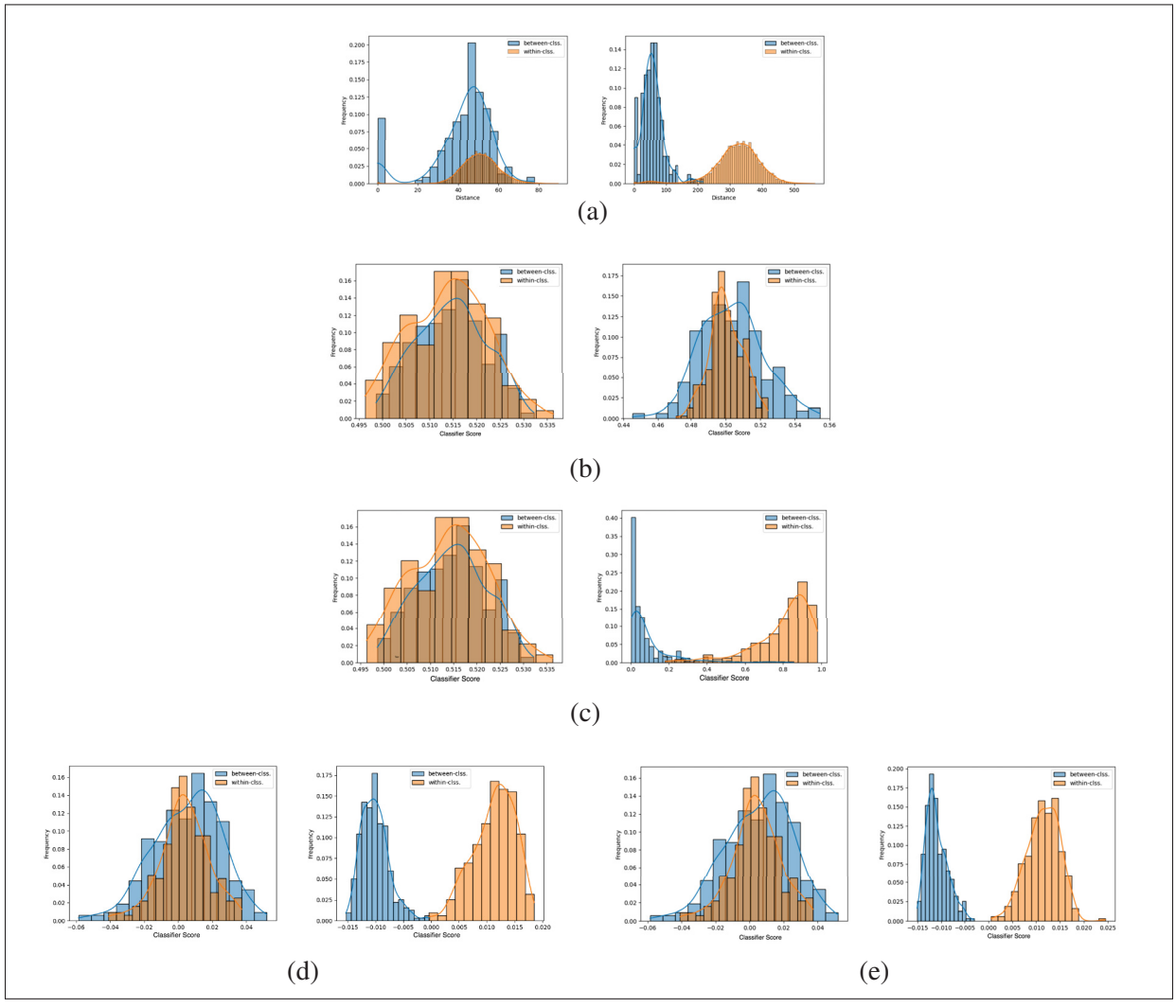

Figure 5.1 A visualization of within- and between-class distance distributions from the Market1501 dataset. The left (right) column indicates distributions before (after) training. (a) Distributions of distances for embedding space; (b) Distribution of classification scores (0 for between and 1 for within class) paired samples in the embedding space, using an MLP trained with cross-entropy; (c) Distribution of classification scores in the dissimilarity space using an MLP trained with cross-entropy (labels 0 and 1); (d) Distribution of classification scores in the dissimilarity space with a max-margin classifier trained (labels: -1 and +1) in a second step with a frozen backbone; (e) Proposed DisReID: end-to-end training of dissimilarity space and max-margin classifier

in the embedding space obtained from a single feature vector by concatenating the embedding pair. Figs 5.1 (c), (d), (e) show the distribution of distance from the separating hyperplane for our proposed dissimilarity-spaced-based classifier trained with cross-entropy loss, max-margin classifier with frozen backbone and end-to-end training with max-margin classifier respectively. Fig. 5.1(e), when dissimilarity space is employed with end-to-end training of the max-margin classifier, their corresponding distances from separating hyperplanes show a clear separation between the classes, reducing the chance of miss-classification. Hence, dissimilarity space is suitable for this problem set.

Since our end application is aimed at real-time performance for ReID with challenges of limited data, we evaluate our proposed method on different lightweight CNN backbones such as mobileNet (Sandler, Howard, Zhu, Zhmoginov & Chen, 2018), ShuffleNet (Zhang, Zhou, Lin & Sun, 2018b), SqueezeNet (Hu, Shen & Sun, 2018a) and OSNet (Zhou *et al.*, 2019). We obtain start-of-the-art results by using a transformer backbone (He *et al.*, 2021) with the DisReID. We evaluate the Market1501 dataset (Zheng *et al.*, 2015b), CUHK03 (Li, Zhao, Xiao & Wang, 2014a) and MSMT17 (Wei, Zhang, Gao & Tian, 2018) dataset to demonstrate the benefit of DisReID in addition to various ablation studies on different components of our proposed method.

Our main contributions are summarized as follows.

(1) An end-to-end training strategy called Dissimilarity ReID (DisReID) is introduced to linearise non-linear high-dimensional features and enable matching with a linear classifier. It relies on a dichotomy transformation to project Euclidean features into dissimilarity space, allowing for joint end-to-end training of a max-margin classifier and the feature extraction backbone.

(2) Dichotomy transformation linearises the feature space, allowing training a classifier with fewer examples per class compared to training in the embedding space.

(3) As opposed to hard sample mining for triple loss, a balanced sampling from a large pool of negative examples is more beneficial for dissimilarity space, avoiding problems with hard sample

mining, such as selection bias and lack of sample diversity (Wu, Manmatha, Smola & Krahenbuhl, 2017b; Xuan *et al.*, 2020; Hermans *et al.*, 2017b).

(4) An extensive set of experiments is presented on the challenging Market1501, CUHK03, and Occluded ReeID datasets, indicating that our proposed DisReID can significantly outperform state-of-art methods using computationally efficient backbone CNNs that are suitable for real-time application. The proposed DisReID has a higher impact on computationally efficient backbones, and in some cases, improvements brought by DisReID on lightweight backbones are comparable to the results of complex backbones.

5.2 Related Work

5.2.1 Person ReID

Person ReID systems are commonly designed by training a backbone network to extract discriminative feature embeddings for similarity matching, which involves optimizing loss functions such as cross-entropy or ID loss (Zheng, Zheng & Yang, 2017c), and contrastive losses like triplet loss (Hermans *et al.*, 2017b). To improve the feature details, fine-grained structures such as part bases model or part alignment with additional information (Suh *et al.*, 2018b), and multi-grain features such as (Zhang *et al.*, 2020c) have been proposed. Recent research has explored vision transformer (ViT) architectures (He *et al.*, 2021; Yu *et al.*, 2022a) and self-supervised learning (Ye *et al.*, 2022; Ma *et al.*, 2023) for person ReID, leveraging unlabelled data. (Yu *et al.*, 2022b) use cascaded transformer architectures to solve the occlusion problem in ReID by progressive refinement of representation. While these methods have improved state-of-the-art accuracy, they are unsuitable for real-time performance.

5.2.2 Efficient ReID architectures

As discussed in the previous sections, various improvements have been proposed to the ReID backbones for robust ReID. However, these improvements come with the additional cost of

computational complexity and are often unsuitable for real-time applications. Efficient person ReID architectures have been proposed to address this challenge. One such architecture is RMNet (Izutov, 2018), designed to be lightweight and computationally efficient for low-power device deployment. The architecture is based on the ResNet (He *et al.*, 2016) framework and incorporates other techniques for constructing efficient convolutional neural networks (Sandler *et al.*, 2018),(Hu *et al.*, 2018a). Through a carefully designed training procedure, RMNet achieves comparable results to heavier ResNet-50-based alternatives, such as MGN(Zhang *et al.*, 2020c). However, fine-grained architectures outperform efficient architectures, but the tradeoff between accuracy and real-time performance is not proportional. Additionally, efficient backbones, such as ShuffleNet (Zhang *et al.*, 2018b) and braidnet (Wang, Chen, Wu & Wang, 2018c), have replaced more complex backbones. However, the tradeoff between accuracy and complexity remains, and few works mentioned above tried to handle this with specific changes in the framework.

Notably, all the discussed methods operate in the embedding space, where a deep neural network maps an input image into a set of floating-point numbers that can be used to match patterns with Euclidean distances or cosine similarity matching. However, using Euclidean space for feature matching in a multi-class problem setting can lead to class overlap. Furthermore, efficient neural networks with fewer parameters may suffer more than deeper, more complex networks. Our work is one of the first to address these problems for the Person ReID application.

5.2.3 Similarity matching in the dissimilarity space

In contrast to the conventional feature space, where each dimension corresponds to a feature value extracted from a single sample, the dissimilarity space entails dissimilarity coordinates. In this space, each dimension represents the disparity between two samples measured concerning a specific feature (Mekhazni *et al.*, 2020).

In the dissimilarity space, a multiclass classification problem can be transformed into a two-class problem (Cha & Srihari, 2000) with dichotomy transformation. Samples from the same classes

are paired together along with the pairing of dissimilar classes. Given two data samples, the problem statement is to identify whether these samples are from the same class rather than which class they belong to. If two samples are from the same class, it is called a positive vector or within-class vector or data. Conversely, the resulting vector is called a negative class if they are from different classes. These vectors can now train a classifier to predict labels in class or between classes. This has been extended to DL by (Souza, Oliveira, Cruz & Sabourin, 2020; Souza *et al.*, 2021), where the dissimilarity transformation has been applied to signature verification applications. Bertolini et al. (Bertolini, Oliveira & Sabourin, 2016) applied dichotomy transformation to convert a k-class problem into a two-class problem for manuscript writer identification. They show that an essential benefit of using dissimilarity space is that the resulting system could identify writers in test time, the examples of which were not used in training. The transformation into a two-class problem generates many samples for the two classes to address a single decision bound. Dichotomy transformation has also been used for bird species identification from audio in (Zottesso, Costa, Bertolini & Oliveira, 2018) where the authors claim that the main advantage of using dichotomy transformation and dissimilarity space is that the classifier trained in dissimilarity space is independent of the number of classes in different test datasets.

Note that the works described above either use handcrafted features or pre-trained deep CNNs with frozen features and only train the classifier. They also indicate that a good representation is necessary for dissimilarity space such that within-class samples cluster closer to the origin and between-class samples are farther away from the origin. Hence, since a CNN with frozen parameters can be suboptimal, we propose a system where the deep backbone network and the classifier can be trained end-to-end in the dissimilarity space to optimize the representations specifically for it.

5.3 Proposed Dissimilarity ReID Method

Inspired by earlier methods in dissimilarity-based learning (Oliveira; Costa *et al.*, 2020a), the proposed method aims to maximize the separation of dissimilar classes by end-to-end

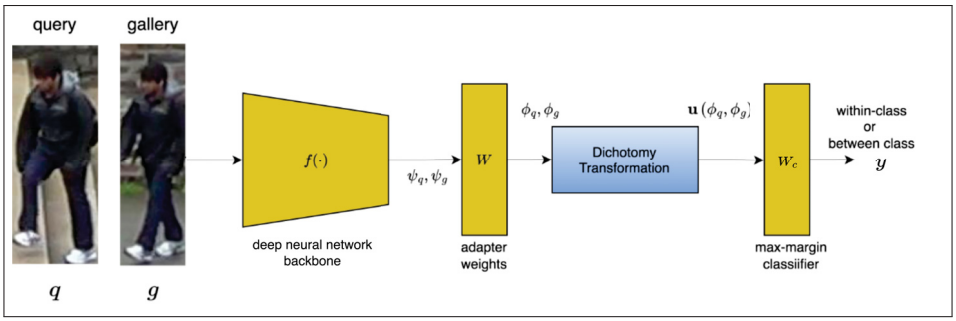

Figure 5.2 The DisReID framework for end-to-end learning of accurate real-time person ReID in the dissimilarity space. A backbone model $f(\cdot)$ extracts feature embeddings from input images that undergo dichotomy transformation to obtain a dissimilarity vector $\mathbf{u}\left(\phi_q, \phi_g\right)$. The max-margin classifier outputs the dissimilarity score and positive or negative class prediction

optimization of the backbone feature extractor with a linear classifier in the dissimilarity space. In addition, a simple, lightweight adapter module is used before transforming the features to dissimilarity space from Euclidean space. Our proposed DisReID framework is illustrated in Fig. 5.2.

The deep neural network backbone $f(\cdot)$ extracts representation vectors from image data samples $\boldsymbol{I}$. The embedding extracted from this is represented by $\psi = f(\boldsymbol{I})$ where ψ is the flattened output with the adaptive average pooling layer. Our architecture allows any deep neural network feature extractors to be used for $f(\cdot)$. For similarity matching, two input images, $\boldsymbol{I_q}$ (query) and $\boldsymbol{I_r}$ (reference), are processed by the backbone, allowing to extract feature vectors ψ_q and ψ_r, respectively.

In the adapter layer, a set of linear weights, one weight per element, is applied to the features ψ before processing by the dichotomizer module. The backbone pretrained network has been trained in the Euclidean space. However, the dissimilarity embedding space can be sparse, particularly the within-class examples close to the origin. Therefore, the distribution is very different from that of Euclidean space. Hence, the adapter layer helps to adapt the Euclidean to the dissimilarity feature embeddings. The adapted features ϕ are defined by $\phi = \sigma\left(W^{(1)}(\psi)\right)$, where W are the weights, and σ is an activation unit (in our case ReLU).

The dichotomy transformation block facilitates the conversion of a multi-class problem into a binary classification task. This process involves organizing the extracted features into pairs representing the absolute differences between the feature vectors, which is then used for binary classification. The dichotomy transformation approach, as introduced by Cha et al. (Cha & Srihari, 2000), achieves this conversion through a two-step procedure. Initially, a combination of input features is created to form pairs. These pairs are categorized as within-class pairs, originating from the same class, or between-class pairs if they differ. Subsequently, the pairs are further transformed into a single feature vector within the dissimilarity space. Let ϕ_q and ϕ_q be query and gallery feature vectors extracted from image samples $\boldsymbol{I}_q$ and $\boldsymbol{I}_g$, respectively. Then, the dissimilarity vector resulting from the dichotomy transformation can be represented by,

$$\mathbf{u}\left(\phi_q, \phi_g\right) = \begin{bmatrix} \left|\varphi_{q1} - \varphi_{g1}\right| \\ \left|\varphi_{q2} - \varphi_{g2}\right| \\ \vdots \\ \left|\varphi_{qn} - \varphi_{gn}\right| \end{bmatrix}, \tag{5.1}$$

Eqn. 1 calculates absolute value of the difference between i_{th} and j_{th} dimension of the embedding, i.e., φ_{qi} and φ_{gi} of the feature vectors ϕ_q and ϕ_g respectively. Dissimilarity space is expected to map a given sample obtained after dichotomy transformation close to the origin of the space if the sample is a within-class sample and farther away from the origin if the sample is a between-class sample. This enables a linear classifier like a max-margin classifier to find an optimal separating hyperplane.

A max-margin classifier $W_c\left(\cdot\right)$ classifies a given dissimilarity vector $\mathbf{u}\left(\phi_q, \phi_g\right)$ into positive or negative classes. It outputs a binary class $\mathbf{c}$ using the dissimilarity feature vector from the previous step, $n\left(\mathbf{u}\right)$.

End to End Training: Given a mini-batch of training data, within-class samples and between-class sample pairs undergo dichotomy transformation after extracting the embeddings with the backbone. The embeddings used for the dichotomy transformation are output embeddings of the adaptive average pooling layer, meaning they are 3D tensors smaller than the feature map output

from the deep neural network and the adapter layer. The backbone deep neural network(DNN) feature extractor may also be trained simultaneously with conventional triplet loss or other contrastive loss functions, as discriminative features are important for the dissimilarity space. The backbone feature extractor training is performed on the output embeddings of the adaptive average pooling layer, CNN's global average pooling layer, and adapter layer to follow the conventional training protocol in person ReID (Ye *et al.*, 2021).

The positive samples represented by y are labeled 1, and negative examples are labeled -1. The max-margin classifier (Cortes & Vapnik, 1995) is essentially a fully connected layer with weights W_c trained with hinge loss. As discussed, the losses are jointly optimized along with L_2 norm on the weights W_c of the linear classifier to train a max-margin classifier,

$$\begin{aligned} \mathcal{L}_{\text{hinge}} &= \tfrac{1}{2}\|\boldsymbol{W_c}\|^2 + \\ & C \sum_{n=0}^{N} \max\left(0, 1 - y_n \left(\boldsymbol{W_c}^{\top} \boldsymbol{u}_n + b\right)\right) \end{aligned} \tag{5.2}$$

The system must be jointly optimized in the embedding and dissimilarity space as a good feature representation is important for dissimilarity space to work in practice (Duin *et al.*, 2010; Oliveira; Costa *et al.*, 2020a). The feature space is optimised by φ with triplet($\mathcal{L}_{\text{tri}}$) loss (Mekhazni *et al.*, 2020) and cross-entropy($\mathcal{L}_{\text{CE}}$) (Mekhazni *et al.*, 2020) for ID classification . Finally, the joint loss is optimized for the convergence of the feature extractor and Max-Margin classifier.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{hinge}} \tag{5.3}$$

Algorithm 2 End-to-end DisReID optimization

Require: batch of image pairs $qg_1, ..., qg_B$, of gallery size B

model parameters θ

Ensure: Binary classification of input pairs y, trained parameters $\theta(\mathrm{f}(\cdot), W_c, W)$

ψ_q , $\psi_g \leftarrow f(q), f(g)$ ▹ Obtain embeddings

ϕ_q , $\phi_g \leftarrow \mathrm{W}(\psi_q), W(\psi_g)$ ▹ Adapted embeddings

$\mathbf{u} \leftarrow \phi_g, \phi_q$ ▹ Convert to dissimilarity vector

$y \leftarrow W_c(\mathbf{u})$ ▹ Binary classification

Compute hinge loss L_{hinge} from Eqn 5.3

Update model parameters: $Optimize(L_{hinge}(\theta))$

The end-to-end optimization of the DisReID framework is summarized in Algorithm 2, where the overall model parameters are represented by θ. The model parameters are updated by optimizing Eqn 5.3.

5.4 Experimental Results and Discussion

Datasets: In this section, we evaluate the proposed method on multiple datasets, including benchmarks such as Market1501 (Zheng *et al.*, 2015b), CUHK03 (Li *et al.*, 2014a), and MSMT17 (Wei *et al.*, 2018). The Market1501 dataset is one of the largest datasets available for person ReID. It contains 1,501 identities and 32,668 gallery images. It also includes 3,368 query images. All these images are captured using six cameras. CUHK03 dataset uses five camera pairs to generate 13,164 images of 1,467 identities. We use the 767/700 split to evaluate the CUHK03 dataset similar to (Saquib Sarfraz *et al.*, 2018). MSMT is a large dataset with over 4,000 identities and 126,000 images. Common ReID performance measures such as Rank-1 and mAP (Ming *et al.*, 2022) have been used in the experiments.

Implementation Details: For validation, different deep neural network-based backbones were used to compare the effectiveness of our proposed method. During training, a pre-trained

network is optimized in the Euclidean space with conventional ReiD losses such as triplet loss and identity loss (Ming *et al.*, 2022). The dissimilarity space is optimized with Eqn. 5.3 to train the max-margin classifier. A learning rate of 0.0003 is used, keeping the batch size 128 and trained with ADAM optimizer.

Inference: During inference of conventional ReID models, gallery representation $(g_1, g_2, ..., g_N)$ of dimension d and Query embedding q of dimension d uses a dot product operation $(q^\intercal[g_1, g_2, ..., g_N])$ for identity matching. opposed to this, our method obtains gallery and query embeddings $(g_1, g_2, ..., g_N)$, q, as usual, and using Eqn.1(main manuscript) to generate dissimilarity features through element-wise subtraction $(gs_1, gs_2, ..., gs_N)$. Identity matching is obtained with dot product between Classifier weights W of dimension d and dissimilarity features $(W^\intercal[gs_1, gs_2, ..., gs_N])$, resulting in classifier scores. Finally, a max operation is performed on the classification scores to determine the identity. Comparing the processes, the only extra operation is the subtraction of gallery and query embeddings, involving $d \times N$ operations(Batchwise operation), which is a few kilo FLOPS. This is minimal compared to saved computations (a few GFLOPS) using lightweight backbones. Please refer to the supplementary material for additional implementation details, details on optimizing the weight norm of the max-margin classifier, and results related to improved training speed with dissimilarity space.

Experimental Setting: The train and test images have been resised to 256x128. Other settings such as ID distribution, number of samples per ID, and train/test splits follow conventional settings as in (Mekhazni *et al.*, 2020). Our model was trained for 120 epochs with a learning rate of 0.003.

5.4.1 Evaluation with lightweight backbones

Our study involved conducting experiments on several computationally efficient convolutional neural network (CNN) backbones by integrating them with our proposed DisReID framework. We refer to the backbone trained with our proposed framework as DisReID, while those trained on the regular CNN embedding space are referred to as "trained in embedding space." In addition to the efficient backbones, we tested our method on ResNet50, a commonly used backbone in

ReID. The computational complexity of each backbone, in terms of the number of floating-point operations (FLOPS), is indicated in Tab. 5.1. Our proposed method consistently improved the Rank-1 accuracy across different backbones. Notably, MobileNetv2 trained with our proposed framework outperformed ResNet50 on the CUHK03 dataset, despite the MobileNetv2 backbone having only one-fourth the computational complexity of ResNet50.

It is worth mentioning that although OSNet trained in the embedding space performed better than MobileNetv2 trained with the DisReID framework on all three datasets, OSNet is twice as complex as MobileNetv2. Furthermore, OSNet trained with the DisReID framework demonstrated an improvement of 1.5-2% over OSNet trained in the embedding space. MobileNetv2 also showed significant improvement, with 3.5% on the Market1501 and 19% on the CUHK03 dataset. A similar trend can be observed with other backbones, and the DisReID framework, in general, resulted in a much greater improvement.

Table 5.1 Rank-1 (R1) accuracy and mAP precision of our proposed DisReID framework (indicated by "ours") with different efficient deep CNN backbones evaluated on ReID datasets. Results on ResNet have been shown for reference. The results are compared against corresponding backbones trained only using the embedding space

Backbone	**Market1501**		**CUHK03**		**MSMT17**		**FLOPS**
	R-1	**mAP**	**R-1**	**mAP**	**R1**	**mAP**	
ShuffleNet	85.1	66.0	39.2	38.9	41.5	19.9	0.15G
ShuffleNet (Ours)	87.9	67.4	44.5	44.1	43.7	22.3	
Improvement	2.8	1.4	5.3	5.2	2.2	2.4	
SqueezeNet	80.1	58.3	40.1	39.7	42.3	18.7	0.4G
SqueezeNet (Ours)	84.1	63.4	47.5	44.6	43.7	23.5	
Improvement	4.1	5.1	7.4	4.9	1.4	5.2	
MobileNetv2	86.0	68.3	47	46.5	51.5	29	0.5G
MobileNetv2 (Ours)	90.5	74	68.9	66	55.1	33.6	
Improvement	4.5	5.7	21.9	19.5	3.6	4.6	
OSNet	94.2	84.9	72.3	67.8	78.7	52.9	0.9G
OSNet (Ours)	94.9	86.2	74.1	68.0	79.2	54.1	
Improvement	0.7	1.3	1.8	0.2	0.5	1.2	
ResNet50	88.2	75.4	60.6	58.4	54.4	37.8	2.7G
ResNet50 (Ours)	91.0	76.0	65.5	62.0	55.3	39.0	
Improvement	2.8	0.6	4.9	3.6	0.9	1.2	

The DisReID framework generates many positive and negative dissimilarity vectors for training by creating pairs of examples from a mini-batch during training. This approach effectively addresses the issue of limited training samples, which ultimately leads to improved results on datasets such as CUHK03. Moreover, in a given mini-batch of samples, although there may only be a few limited examples for each class, the combination of every other class in the dissimilarity space serves as a negative pair. As noted in (Chen, Kornblith, Norouzi & Hinton, 2020c), many negative examples can significantly enhance the overall representation power of convolutional neural networks (CNNs). Contrastive losses, such as triplet loss, face challenges in achieving this due to their heavy reliance on hard sample mining (Hermans *et al.*, 2017b; Boudiaf *et al.*, 2020b). Training the max-margin classifier with the hinge loss solves some of the problems with sample mining as the system selects the relevant samples with the help of hinge loss. The max-margin classifier (SVM) addresses the problem of hard sample mining by directly incorporating the support vectors, which are the most challenging instances that lie closest to the decision boundary. By maximizing the margin and considering these informative samples, the SVM algorithm naturally focuses on the difficult instances, improving the classification performance on hard samples without having to mine hard samples explicitly, as in Triplet loss.

5.4.2 Comparison with state-of-the-art ReID

Tab. 5.2 shows the results of comparing our proposed method with some of the state-of-the-art ReID methods. We implement the TransReID (He *et al.*, 2021) vision transformer backbone into our DisReID framework to obtain state-of-the-art results in Market1501, CUHK03, and MSMT17 datasets. The results of Tab. 5.2 and Tab 5.1 show that our method is agnostic to the backbone used and can be robust in different application settings. As previously discussed, we attribute the effectiveness of the result to the dissimilarity space and max-margin classifier, which solves some of the problems with Euclidean space-based distance matching. We can observe that we obtained state-of-the-art results in cukh03 and the MSMT17 dataset on Rank-1 accuracy and mAP. On the market dataset, although the results considering the Rank-1 accuracy

are lower than that of MGN (Wang, Yuan, Chen, Li & Zhou, 2018a), we obtain improved results on the mAP.

Table 5.2 Rank-1 (R1) accuracy and mAP precision of our proposed DisReID framework (indicated by "ours") implemented with TransReID backbone. Our proposed method is compared with some of the state-of-the-art ReID methods

Method	**Venue**	**Backbone**	Market1501		cukh03		MSMT17	
			R-1	mAP	R-1	mAP	R-1	mAP
Computationally Efficient Backbones								
MobileNetV2 (Sandler *et al.*, 2018)	CVPR18	MobileNetV2	86	68.3	47	46.5	51.5	29
OSNet (Zhou *et al.*, 2019)	ICCV19	OSNet	94.8	84.9	72.3	67.8	78.7	52.9
DisReID(Ours)		MobileNetV2	90.5	74	68.9	66	55.1	33.6
DIsReID(Ours)		OSNet	94.9	86.2	74.1	68	79.2	54.1
Computationally Complex Backbones								
MGN (Wang *et al.*, 2018a)	ACM18	Modified ResNet50	95.7	86.9	-	-	76.9	52.1
PCB (Suh *et al.*, 2018a)	ECCV18	ResNet50+Partial Feat.	93.8	81.6	63.7	57.5	-	-
DGNet (Zheng *et al.*, 2019d)	CVPR19	ResNet50+Decoder	94.8	86	-	-	77.2	52.3
SAN (Jin, Lan, Zeng, Wei & Chen, 2020)	AAAI20	ResNet50+Decoder	88.4	96.1	-	-	79.2	55.7
TransReID (He *et al.*, 2021)	ICCV21	VisionTransformer	95.2	89.5	-	-	86.2	69.4
Nformer (Wang, Shen, Liu, Gao & Gavves, 2022a)	CVPR22	ResNet50	94.7	91.1	77.2	78	77.3	59.8
BPBreID (Somers, De Vleeschouwer & Alahi, 2023a)	WACV23	HRNet	95.7	89.4	76.5	77.1	75.9	56.4
PHA (Zhang, Zhang, Zhang, Li & Pu, 2023a)	CVPR23	Transformer	96.1	90.2	84.5	83	86.1	68.9
DisReID(Ours)		VisionTransformer	95.5	90.2	74.5	68.9	88.9	72.3
DisReID(Ours)		HRNet	95.6	91.4	77.7	78.3	79.8	64.7

5.4.3 Ablation study.

(a) Different components of our proposed system:

We further study the effects of different components of our proposed system. The results from Tab. 5.3 were generated using the mobileNetv2 backbone in our DisReId framework on Market1501 and CUHK03 datasets. The study started with dichotomy transformation and introduced a max-margin classifier and pre-trained but frozen backbone, as shown in row 2. The backbone was pre-trained in the embedding space using the corresponding dataset in the respective columns of the Table. Row 2 shows that the max-margin classifier and dichotomy transformation can improve the network's results compared to the embedding-based backbone training, as shown in row 1. We perform another experiment where we perform end-to-end training but keep the components the same as row 2, and the results of this further show the benefit of joint training. Another experiment was performed to test the necessity of a max-margin classifier. An MLP replaced the max-margin classifier with cross-entropy loss to classify a dissimilarity vector as in-class or between-class. This is shown in row 4, and the result

shows that the max-margin benefits the dissimilarity setting. This is because, after dichotomy transformation, the data becomes linearly separable, and max-margin optimization is ideal for such a setting. Finally, the feature adapter module is applied for end-to-end training, where the overall results improved by a big margin of nearly 10% on the CUHK03 dataset. The results from Tab 5.3 can be visualized in Fig. 5.1 where an attempt to visualize classification performance has been made in the form of a histogram of the distribution of within-class distances for Euclidean space and in the form of a histogram of classification scores for the proposed DisReID framework (shown in Fig. 5.1(d). In Fig.5.1(b), the distribution of classification scores for within-class and between-class samples in the dissimilarity space is presented, along with a binary classifier trained using binary cross-entropy loss. This training significantly improves the separation; however, the borders still exhibit marginal overlap. Fig.5.1(c) demonstrates the distribution of classification scores in the dissimilarity space when utilizing a margin-based classifier with a frozen backbone feature extractor. Compared with Fig.5.1(a), the dissimilarity space already improves the distribution of classification scores compared to the distribution of distances in Euclidean space.

Table 5.3 Ablation study showing the impact of different DisReID components used during the training and the classifiers used in the dissimilarity space

DisReID Components			**Market1501**		**CUHK03**	
Adapter	**Classifier**	**Training**	**R-1**	**mAP**	**R-1**	**mAP**
×	×	×	86.0	68.3	47.0	46.5
×	max-margin	frozen	88.0	72.4	49.4	48.6
×	max-margin	joint	89.1	73.2	53.6	49.1
✓	MLP	joint	89.5	73.1	58.0	54.4
✓	max-margin	joint	90.5	74.0	68.9	66.0

(b) Between-class sampling strategy:

During training, in-class and between-class samples are generated using dichotomy transformation from a mini-batch of multi-class examples. However, this results in many between-class samples or samples with negative labels, which leads to data imbalance during training. To address this issue, we propose different strategies. The first strategy, “Balanced ranking," involves ranking

the negative examples based on their Euclidean distances from positive samples and selecting the top samples so that their total count matches the positive ones. Ranking by distances is similar to the strategy used with triplet loss (Hermans *et al.*, 2017b; Boudiaf *et al.*, 2020b). The second strategy, called “All Samples," involves using all the negative examples during dichotomy transformation, resulting in data imbalance. The third strategy, “Balanced random," involves producing all possible between-class samples and randomly selecting samples to match the count of in-class samples for data balancing.

The results presented in Table 5.4 indicate that the "Balanced random" sampling strategy outperforms the “Balanced rank" strategy, or using all the samples. This can be attributed to the fact that selecting samples based on their Euclidean distance from the positive class may not represent their distance in a non-linear embedding space. In contrast, using all the samples from between-class pairs leads to data imbalance during training, adversely affecting the results. But “Balanced random" sampling avoids problems with hard sample mining, like an amplification of label noise and lack of diversity, which triplet loss suffers from. At the same time, the system takes care of the selection of hard samples, especially with the max-margin formulation of the classifier.

Table 5.4 Impact on DisReID accuracy of the strategy used for between-class sampling

Between-Class Sampling	**Market1501**		**CUHK03**	
	R-1	**mAP**	**R-1**	**mAP**
balanced ranked	87.5	71.2	59.1	56.7
all samples	88.3	73.6	62.6	59.8
balanced random	90.5	74.0	68.9	66.0

(c) Effect of the number of training samples and classes.

This study aims to compare the performance of deep networks trained on embedding space and DisReID framework and evaluate the results using various fractions of the Market1501 dataset and ResNet50 backbone. The findings are presented in Tab 5.5 and Tab 5.6. Two different sets of experiments were conducted. Tab 5.5 shows the experiment where fewer examples are used

per class for training and the percentage of examples used is varied, keeping the number of classes the same as in the original dataset. Tab 5.6 is a similar experiment where the number of classes is reduced and the percentage of classes used for training is varied. The results indicate that the Rank-1 accuracy of the network trained with the DisReID framework on half the dataset in the dissimilarity space is comparable to that of the model trained in the embedding space with the full dataset. In particular, the mapping score is improved and, in some cases, up to 5% (improvements shown by the columns labeled Δ). The difference in mAP is highlighted with Δ as mAP has a much higher impact than Rank-1. Higher mAP also refers to consistent duplicates of the retrieved persons from the gallery.

DisReID framework-based training transforms the problem into a two-class problem, ensuring that the ratio of examples to the number of classes (two classes) is relatively higher than the multi-class problem. In addition, the dissimilarity space projection has a much smaller overlap between in-class and between-class samples, making the learning problem much easier. Therefore, DisReID has an advantage when training with fewer examples per class.

Table 5.5 Impact of training in the dissimilarity space with fewer classes (reduction in the number of IDs/classes) in the training data

Percentage of Classes	**Euclidean**		**Dissimilarity**		**Δ mAP**
	mAP	**R-1**	**mAP**	**R-1**	
100%	75.4	88.2	76.0	91.0	0.6
70%	64.9	83	69.8	85.2	4.9
50%	60.2	79.5	65.4	81.8	5.2
30%	51.9	73.5	58.1	76.5	6.2

5.5 Conclusion

In this work, we propose DisReID, an end-to-end learning framework for real-time Person ReID applications that utilize the dissimilarity space and a max-margin classifier. Our findings demonstrate that our DisReID can boost the performance of efficient CNN backbones and improve the tradeoff between computational efficiency and accuracy. Unlike commonly used

Table 5.6 Study on training the dissimilarity space with the reduced number of samples per ID/class

Percentage of dataset	**Euclidean**		**Dissimilarity**		**Δ mAP**
	mAP	**R-1**	**mAP**	**R-1**	
100%	75.4	88.2	76.0	91.0	0.6
70%	71.3	86.2	74.8	8.3	3.5
50%	65.7	82.5	69.6	84.1	3.9
30%	59.7	78.6	64.5	81.5	4.8

embedding spaces for ReID, our approach involves end-to-end learning of a max-margin classifier in the dissimilarity space. It uses classification scores of pair-wise samples for matching. DisReID framework has a higher impact with efficient, lightweight architectures improving the performance of ReID in real-time applications. DisReID framework with efficient backbones and trained with fewer training examples can train faster and perform on par with computationally expensive backbones trained with a much larger number of examples. We also achieve competitive performances with DisReID trained with transformers backbone. The results suggest that dissimilarity space may be a viable alternative to the embedding space for metric learning-like problems.

CONCLUSION AND RECOMMENDATIONS

6.1 Summary of Contributions

In light of the challenges and limitations discussed earlier, the principal inquiry of this research revolves around the question: *Can a proficient video representation be acquired to address the intricacies of tracking and ReID, conceptualizing them as a matching problem while simultaneously navigating the challenges posed by occlusion, the intricacies of video representation learning, and the computational complexities inherent in feature extraction?* This study is intricately woven around the exploration of three distinct research trajectories, emphasizing adaptive tracking for precise object localization and bolstering the resilience of object and person re-identification in real-time video surveillance.

Chapter 2 of our thesis focuses on improving online learning in visual object tracking (VOT) models, particularly addressing the challenges of concept drift. Our study found that a gradual concept drift helps adapt to changing target appearances, while abrupt changes often result from distractions like occlusion, requiring careful avoidance of model updates. Additionally, recurring drifts happen when the object's previous appearance reoccurs, and maintaining a sample buffer with high variance proves effective for online tracking.

To tackle the issue of occlusion causing drift in object tracking, we proposed an occlusion-aware training approach to focus on the visible regions of the object. We applied this method to enhance the DiMP, PrDiMP, and SuperDiMP trackers, demonstrating a significant 2% improvement in Area Under the Curve (AUC) when evaluated on various datasets like OTB, LaSOT, TrackingNet, and UAV. Importantly, our method also showcased a higher frame rate, a crucial factor for video surveillance applications. By maintaining an optimal number of samples, our system reduces the chance of noise introduction by adding too many samples, and the memory is optimized, which is once again crucial for real-time video surveillance systems. By performing online classifier

training only when a gradual change is detected, our tracker achieved a faster frame rate, making it more suitable for real-world video surveillance applications.

Chapter 3 discusses the domain of video person ReID within video analytics and surveillance applications. This chapter has delved into addressing this limitation by exploring the motion pattern of individuals as an additional cue for ReID. The proposed approach introduces a Flow-Guided Mutual Attention network, which adeptly fuses bounding box and optical flow sequences over tracklets, utilizing a 2D-CNN backbone to encode temporal and spatial appearance information. Additionally, a novel method for aggregating features from extended input streams has been introduced to augment video sequence-level representation. Experimental results indicate a significant improvement in recognition accuracy compared to conventional gated-attention networks and state-of-the-art methods in video-based person ReID. This work underscores the potential of motion-pattern-guided attention mechanisms in advancing the capabilities of deep learning models for robust video ReID applications.

Chapter 4 introduces a novel holistic-generative (HG) student-teacher model tailored for occluded person ReID, which dispenses with the need for image identity labels and circumvents resource-intensive processes that focus solely on visible segments of occluded regions. The proposed HG teacher leverages the Distribution of Class Distances (DCD) across samples in a comprehensive dataset to train a student model, enabling the generation of attention maps and mitigating the challenges posed by occlusion. Diverging from prevalent approaches in the literature that incorporate external supervision, such as pose, to create visibility cues, our method relies solely on the holistic data distribution during training, treating it as a soft label. Consequently, during testing, our model operates seamlessly without the requirement for external cues like pose, with the overall parameters comprising solely the backbone Encoder and a compact embedding for attention map generation during feature extraction. Additionally, joint learning of a denoising autoencoder has been employed to enhance the model's capacity

for self-recovery from occlusion. Empirical evaluations on diverse and challenging datasets demonstrate the superior performance of our HG methodology, surpassing state-of-the-art models in Occluded-ReID and Holistic ReID tasks.

Chapter 5 tackles the problem of the scarcity of training data and its effect on the matching accuracy for ReID problems, along with the limitations imposed by deploying lightweight architectures. Fine-tuning these lightweight structures on limited annotated data is recognized as causing issues such as the overlapping and dispersal of class distributions within the feature space. The conventional approach of employing Euclidean feature space mapping in standard ReID methods is problematic, mainly when dealing with high-dimensional manifolds, class imbalance, and overlap instances. By addressing issues related to limited training data, lightweight architectures, and challenges in the Euclidean feature space, DisReID introduces the concept of dichotomy transformation to project features into a dissimilarity space. The proposed end-to-end training approach optimizes the backbone and max-margin classifier simultaneously, focusing on achieving linear separation of within- and between-class samples. The effectiveness of DisReID suggests promising advancements in overcoming critical obstacles faced by standard ReID methods, offering potential applications in real-time video surveillance scenarios.

6.2 Limitations and Recommendations

Limitations and Recommendation

Following are some of the limitations of our work in this thesis:

- **Use of standard backbone feature extractors** All our work is based on standard backbone CNN feature extractors like ResNet (He *et al.*, 2016). However, it is important to note that although this is common in other related works for tracking and ReID, ResNet was originally designed for object classification ImageNet challenge Russakovsky *et al.* (2015a). Object classification often requires spatial invariance in feature representation. However, ReID

and tracking are very dependent on spatial information. For example, some of the person examples in ReID can have subtle differences. Hence, ResNet might not be able to extract representative embeddings.

- **Concept drift detection mechanism could be weak** The main reason for including a simple change detection mechanism is computational efficiency. We introduce two methods: one uses MMD distance for distribution-based drift detection, and the other uses statistics of the classifier scores. This could be further improved by using a trained model to model time-series data, such as RNNs or LSTMs, and take advantage of the non-linear representation of time-series data with neural networks and RNNs.
- **Mutual attention does not address synchronization issues** We propose to use Mutual attention between optical flow and Deep CNN features for video person ReID. However, this might not be ideal as optical flow is calculated between two different frames, but it is used to attend to features extracted from deep CNN. It is possible that the optical flow and deep CNN features are not time synchronized, i.e., they might differ by at least one frame. This might not efficiently take advantage of both modalities.
- **Simple distance function used for dissimilarity space transformation**
  An element-wise distance calculation on Deep CNN features achieved the dissimilarity space representation. Exploring other distance functions, including non-linear distances, could further improve this.

Based on the above limitations of our work, we propose the following recommendations,

- **Refinement of Concept Drift Handling**: Explore methods to refine how visual object tracking models handle different types of concept drift, such as gradual, abrupt, and recurring changes. This could involve developing more nuanced strategies for adapting to each type of drift. In particular, use a neural network or learning-based method to detect drift. Take advantage of available training data in tracking datasets to learn concept drift.

- **Enhancement of Classifier Confidence**: Focus on enhancing the reliability of classifier confidence scores for detecting track drift. Explore and implement calibration methods to ensure more accurate and trustworthy confidence scores, particularly in practical scenarios with specific limitations.
- **Explore application-specific deep CNN architecture**: Deep CNN architecture for tracking and ReID needs to focus more on spatial information than representation power. This could already solve some of the challenges in both applications.

BIBLIOGRAPHY

Zelkowitz, M. V. (Ed.). (2010). Advances in Video-Based Human Activity Analysis: Challenges and Approaches. *Advances in Computers* (vol. 80, pp. 237 - 290). Elsevier.

Abadi, M. & et al. [Software available from tensorflow.org]. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from: https://www.tensorflow.org/.

Agrahari, S. & Singh, A. K. (2021). Concept Drift Detection in Data Stream Mining: A literature review. *Journal of King Saud University-Computer and Information Sciences*.

Ahmed, E., Jones, M. & Marks, T. (2015). An improved deep learning architecture for person re-identification. *CVPR*.

Aljundi, R., Lin, M., Goujaud, B. & Bengio, Y. (2019a). Gradient based sample selection for online continual learning. *arXiv:1903.08671*.

Aljundi, R., Lin, M., Goujaud, B. & Bengio, Y. (2019b). Gradient based sample selection for online continual learning. *NIPS*.

Avidan, S. (2007). Ensemble Tracking. *IEEE Trans. PAMI*, 29(2), 261-271.

Ba, J. & Frey, B. (2013). Adaptive dropout for training deep neural networks. *NIPS*, pp. 3084–3092.

Baharani, M., Mohan, S. & Tabkhi, H. (2019). Real-time person re-identification at the edge: A mixed precision approach. *International Conference on Image Analysis and Recognition*, pp. 27–39.

Barros, R. S., Cabral, D. R., Gonçalves Jr, P. M. & Santos, S. G. (2017). RDDM: Reactive drift detection method. *Expert Systems with Applications*, 90, 344–355.

Bengio, E., Bacon, P.-L., Pineau, J. & Precup, D. (2015). Conditional computation in neural networks for faster models. *arXiv preprint arXiv:1511.06297*.

Bengio, Y., Léonard, N. & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Bergmann, P., Meinhardt, T. & Leal-Taixe, L. (2019, October). Tracking Without Bells and Whistles. *The IEEE International Conference on Computer Vision (ICCV)*.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A. & Torr, P. H. (2016). Fully-Convolutional Siamese Networks for Object Tracking. *arXiv preprint arXiv:1606.09549*.

Bertolini, D., Oliveira, L. S. & Sabourin, R. (2016). Multi-script writer identification using dissimilarity. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 3025–3030.

Bewley, A., Ge, Z., Ott, L., Ramos, F. & Upcroft, B. (2016a). Simple online and realtime tracking. *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468.

Bewley, A., Ge, Z., Ott, L., Ramos, F. & Upcroft, B. (2016b). Simple online and realtime tracking. *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468.

Bhat, G., Danelljan, M., Gool, L. V. & Timofte, R. (2019a). Learning discriminative model prediction for tracking. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6182–6191.

Bhat, G., Danelljan, M., Gool, L. V. & Timofte, R. (2019b). Learning discriminative model prediction for tracking. *CVPR*.

Bhat, G., Danelljan, M., Van Gool, L. & Timofte, R. (2020). Know your surroundings: Exploiting scene information for object tracking. *ECCV*, pp. 205–221.

Bhuiyan, A., Perina, A. & Murino, V. (2014). Person re-identification by discriminatively selecting parts and features. *ECCV*.

Bhuiyan, A., Perina, A. & Murino, V. (2018). Exploiting Multiple Detections for Person Re-Identification. *Journal of Imaging*, 4(2), 28.

Bhuiyan, A., Liu, Y., Siva, P., Javan, M., Ayed, I. B. & Granger, E. (2020). Pose Guided Gated Fusion for Person Re-identification. *WACV*.

Bifet, A. & Gavalda, R. (2007). Learning from time-changing data with adaptive windowing. *SIAM*.

Bolme, D. S., Beveridge, J. R., Draper, B. A. & Lui, Y. M. (2010). Visual object tracking using adaptive correlation filters. *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 2544–2550.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P. & Ayed, I. B. (2020a). A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, pp. 548–564.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P. & Ayed, I. B. (2020b). A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. *European conference on computer vision*, pp. 548–564.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1993). Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1994). Signature verification using a" siamese" time delay neural network. *NIPS*.

Cai, H., Wang, Z. & Cheng, J. (2019a). Multi-scale body-part mask guided attention for person re-identification. *CVPR Workshops*.

Cai, H., Wang, Z. & Cheng, J. (2019b). Multi-scale body-part mask guided attention for person re-identification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0.

Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*.

Caron, M., Bojanowski, P., Joulin, A. & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149.

Cha, S.-H. & Srihari, S. N. (2000). Writer identification: statistical analysis and dichotomizer. *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, pp. 123–132.

Chang, X., Hospedales, T. M. & Xiang, T. (2018). Multi-level factorisation net for person re-identification. *CVPR*.

Chaudhry, A., Ranzato, M., Rohrbach, M. & Elhoseiny, M. (2018). Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B., Gan, W., Wu, W. & Ouyang, W. (2022a). Backbone is all your need: A simplified architecture for visual object tracking. *European Conference on Computer Vision*, pp. 375–392.

Chen, D., Li, H., Xiao, T., Yi, S. & Wang, X. (2018a, June). Video Person Re-identification with Competitive Snippet-Similarity Aggregation and Co-attentive Snippet Embedding. *ICCV*, pp. 1169-1178. doi: 10.1109/CVPR.2018.00128.

Chen, G., Lu, J., Yang, M. & Zhou, J. (2019). Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 28(9), 4192–4205.

Chen, G., Lu, J., Yang, M. & Zhou, J. (2020a). Learning Recurrent 3D Attention for Video-Based Person Re-identification. *IEEE Transactions on Image Processing*.

Chen, G., Lu, J., Yang, M. & Zhou, J. (2020b). Learning Recurrent 3D Attention for Video-Based Person Re-identification. *IEEE Transactions on Image Processing*.

Chen, L., Ai, H., Zhuang, Z. & Shang, C. (2018b, July). Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6. doi: 10.1109/ICME.2018.8486597.

Chen, L., Ai, H., Zhuang, Z. & Shang, C. (2018a). Real-time multiple people tracking with deeply learned candidate selection and person re-identification. *2018 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020c). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pp. 1597–1607.

Chen, W., Chen, X., Zhang, J. & Huang, K. (2017a). Beyond triplet loss: a deep quadruplet network for person re-identification. *CVPR*.

Chen, X., Peng, H., Wang, D., Lu, H. & Hu, H. (2023). Seqtrack: Sequence to sequence learning for visual object tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14572–14581.

Chen, Y., Zhu, X. & Gong, S. (2017b). Person re-identification by deep learning multi-scale representations. *ICCV*.

Chen, Y., Xia, S., Zhao, J., Zhou, Y., Niu, Q., Yao, R., Zhu, D. & Liu, D. (2022b). ResT-ReID: Transformer block-based residual learning for person re-identification. *Pattern Recognition Letters*, 157, 90–96.

Chen, Z., Badrinarayanan, V., Lee, C.-Y. & Rabinovich, A. (2018b). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *International Conference on Machine Learning*, pp. 794–803.

Cheng, D., Gong, Y., Zhou, S., Wang, J. & Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. *CVPR*.

Cheng, L., Wang, J., Gong, Y. & Hou, Q. (2015). Robust deep auto-encoder for occluded face recognition. *ACMM*.

Chicco, D. (2021). Siamese neural networks: An overview. *Artificial neural networks*, 73–94.

Cho, K., Van Merriënboer, B., Bahdanau, D. & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, S. & Foroosh, H. (2018). Spatio-temporal fusion networks for action recognition. *ACCV*, pp. 347–364.

Cho, Y.-J. & Yoon, K.-J. (2016). Improving person re-identification via pose-aware multi-shot matching. *CVPR*.

Choi, J., Kwon, J. & Lee, K. M. (2018). Real-time visual tracking by deep reinforced decision making. *Computer Vision and Image Understanding*, 171, 10–19.

Choi, J., Kwon, J. & Lee, K. M. (2019). Deep meta learning for real-time target-aware visual tracking. *ICCV*.

Choi, S., Lee, J., Lee, Y. & Hauptmann, A. (2020). Robust Long-Term Object Tracking via Improved Discriminative Model Prediction. *ECCV*, pp. 602–617.

Chopra, S., Hadsell, R. & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 1, 539–546.

Chung, D., Tahboub, K. & Delp, E. J. (2017, Oct). A Two Stream Siamese Convolutional Neural Network for Person Re-identification. *ICCV*, pp. 1992-2000. doi: 10.1109/ICCV.2017.218.

Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R. & Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381, 61–88.

Comaschi, F., Stuijk, S., Basten, T. & Corporaal, H. Online multi-face detection and tracking using detector confidence and structured SVMs. *AVSS 2015*.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20, 273–297.

Costa, Y. M., Bertolini, D., Britto, A. S., Cavalcanti, G. D. & Oliveira, L. E. (2020a). The dissimilarity approach: a review. *Artificial Intelligence Review*, 53, 2783–2808.

Costa, Y., Bertolini, D., Britto, A. S., Cavalcanti, G. D. & Oliveira, L. E. (2020b). The dissimilarity approach: a review. *Artificial Intelligence Review*, 53(4), 2783–2808.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1), 53–65.

Dai, K., Zhang, Y., Wang, D., Li, J., Lu, H. & Yang, X. (2020). High-performance long-term tracking with meta-updater. *CVPR*.

Dai, Z., Yang, Z., Yang, F., Cohen, W. W. & Salakhutdinov, R. R. (2017). Good semi-supervised learning that requires a bad gan. *Advances in neural information processing systems*, 30.

Danelljan, M., Shahbaz Khan, F., Felsberg, M. & Van de Weijer, J. (2014). Adaptive color attributes for real-time visual tracking. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1090–1097.

Danelljan, M., Hager, G., Shahbaz Khan, F. & Felsberg, M. (2016). Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1430–1438.

Danelljan, M., Bhat, G., Shahbaz Khan, F. & Felsberg, M. (2017). Eco: Efficient convolution operators for tracking. *CVPR*.

Danelljan, M., Bhat, G., Khan, F. S. & Felsberg, M. (2019). Atom: Accurate tracking by overlap maximization. *CVPR*.

Danelljan, M., Gool, L. V. & Timofte, R. (2020). Probabilistic regression for visual tracking. *CVPR*.

Dasu, T., Krishnan, S., Venkatasubramanian, S. & Yi, K. (2006). An information-theoretic approach to detecting changes in multi-dimensional data streams. *In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications*.

Davila, D., Du, D., Lewis, B., Funk, C., Van Pelt, J., Collins, R., Corona, K., Brown, M., McCloskey, S., Hoogs, A. et al. (2023). Mevid: Multi-view extended videos with identities for video person re-identification. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1634–1643.

de Andrade Silva, J., Hruschka, E. R. & Gama, J. (2017). An evolutionary algorithm for clustering data streams with a variable number of clusters. *Expert Systems with Applications*, 67, 228–238.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K. & Leal-Taixé, L. (2019). CVPR19 Tracking and Detection Challenge: How crowded can it get? *arXiv:1906.04567 [cs]*. arXiv: 1906.04567.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K. & Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*.

Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H. & Mei, T. (2021). MINet: Meta-Learning Instance Identifiers for Video Object Detection. *IEEE Transactions on Image Processing*.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y. & Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 994–1003.

Diba, A., Pazandeh, A. M. & Van Gool, L. (2016). Efficient two-stream motion and appearance 3d cnns for video classification. *arXiv preprint arXiv:1608.08851*.

Ditzler, G. & Polikar, R. (2011). Hellinger distance based drift detection for nonstationary environments. *2011 IEEE symposium on computational intelligence in dynamic and uncertain environments (CIDUE)*, pp. 41–48.

Dong, S., Xia, Y. & Peng, T. (2021). Network abnormal traffic detection model based on semi-supervised deep reinforcement learning. *IEEE Transactions on Network and Service Management*, 18(4), 4197–4212.

Dong, X. & Shen, J. (2018). Triplet loss in siamese network for object tracking. *ECCV*.

Duin, R. P., Bicego, M., Orozco-Alzate, M., Kim, S.-W. & Loog, M. (2014). Metric learning in dissimilarity space for improved nearest neighbor performance. *IAPR Workshop*.

Duin, R. P. & Pkalska, E. (2011). The dissimilarity representation for structural pattern recognition. *Iberoamerican Congress on Pattern Recognition*, pp. 1–24.

Duin, R. P., Loog, M., Pkalska, E. & Tax, D. M. (2010). Feature-based dissimilarity space classification. *Recognizing Patterns in Signals, Speech, Images and Videos: ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports*, pp. 46–55.

Duman, E. & Erdem, O. A. (2019). Anomaly detection in videos using optical flow and convolutional autoencoder. *IEEE Access*, 7, 183914–183923.

Ekladious, G., Lemoine, H., Granger, E., Kamali, K. & Moudache, S. (2020). Dual-triplet metric learning for unsupervised domain adaptation in video face recognition. *IJCNN*.

Elhamifar, E., Sapiro, G. & Vidal, R. (2012). Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. *NeurIPS*.

Eom, C., Lee, G., Lee, J. & Ham, B. (2021). Video-based person re-identification with spatial and temporal memory networks. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12036–12045.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), 303–338.

Fan, H. & Ling, H. (2019). Siamese cascaded region proposal networks for real-time visual tracking. *CVPR*.

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C. & Ling, H. (2019). Lasot: A high-quality benchmark for large-scale single object tracking. *CVPR*.

Farajtabar, M., Azizan, N., Mott, A. & Li, A. (2020). Orthogonal gradient descent for continual learning. *International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773.

Farenzena, M., Bazzani, L., Perina, A., Murino, V. & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. *CVPR*.

Fischer, A. & Igel, C. (2012). An introduction to restricted Boltzmann machines. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17*, pp. 14–36.

Frias-Blanco, I., del Campo-Ávila, J., Ramos-Jimenez, G., Morales-Bueno, R., Ortiz-Diaz, A. & Caballero-Mota, Y. (2014). Online and non-parametric drift detection methods based on Hoeffding's bounds. *IEEE Transactions on Knowledge and Data Engineering*, 27(3), 810–823.

Fu, Y., Wang, X., Wei, Y. & Huang, T. (2019). Sta: Spatial-temporal attention for large-scale video-based person re-identification. *AAAI 2019*, 33, 8287–8294.

Fu, Z., Liu, Q., Fu, Z. & Wang, Y. (2021). STMTrack: Template-free Visual Tracking with Space-time Memory Networks. *CVPR*.

Gama, J., Medas, P., Castillo, G. & Rodrigues, P. (2004). Learning with drift detection. *Brazilian symposium on artificial intelligence*, pp. 286–295.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4), 1–37.

Gao, J. & Nevatia, R. (2018a). Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*.

Gao, J. & Nevatia, R. (2018b). Revisiting Temporal Modeling for Video-based Person ReID. *arXiv preprint arXiv:1805.02104*.

Gao, S., Wang, J., Lu, H. & Liu, Z. (2020a). Pose-guided Visible Part Matching for Occluded Person ReID. *CVPR*.

Gao, S., Wang, J., Lu, H. & Liu, Z. (2020b). Pose-guided visible part matching for occluded person reid. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11744–11752.

Gao, Y., Beijbom, O., Zhang, N. & Darrell, T. (2016). Compact bilinear pooling. *CVPR*.

Gautam, A. & Singh, S. (2019). Trends in Video Object Tracking in Surveillance: A Survey. *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 729-733. doi: 10.1109/I-SMAC47947.2019.9032529.

Gavves, E., Tao, R., Gupta, D. K. & Smeulders, A. W. (2021a). Model decay in long-term tracking. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2685–2692.

Gavves, E., Tao, R., Gupta, D. K. & Smeulders, A. W. (2021b). Model decay in long-term tracking. *ICPR 2020*, pp. 2685–2692.

Ge, W. (2018a). Deep metric learning with hierarchical triplet loss. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–285.

Ge, W. (2018b). Deep metric learning with hierarchical triplet loss. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–285.

Geng, M., Wang, Y., Xiang, T. & Tian, Y. (2016). Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT press.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. & Smola, A. (2012). A kernel two-sample test. *JMLR*, 13(1), 723–773.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. et al. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, 77, 354–377.

Gu, X., Ma, B., Chang, H., Shan, S. & Chen, X. (2019). Temporal Knowledge Propagation for Image-to-Video Person Re-identification. *ICCV*.

Guo, D., Wang, J., Cui, Y., Wang, Z. & Chen, S. (2020). SiamCAR: Siamese fully convolutional classification and regression for visual tracking. *CVPR*.

Guo, J., Xu, T., Jiang, S. & Shen, Z. (2018). Generating reliable online adaptive templates for visual tracking. *ICIP 2018*, pp. 226–230.

Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L. & Wang, S. (2017). Learning dynamic siamese network for visual object tracking. *ICCV*, pp. 1763–1771.

Gustafsson, F. (2000). *Adaptive Filtering and Change Detection*. Chichester, United Kingdom: John Wiley and Sons Ltd.

Hadash, G., Kermany, E., Carmeli, B., Lavi, O., Kour, G. & Jacovi, A. (2018). Estimate and Replace: A Novel Approach to Integrating Deep Neural Networks with Existing Applications. *arXiv preprint arXiv:1804.09028*.

Hadsell, R., Chopra, S. & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2, 1735–1742.

Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M. M., Hicks, S. L. & Torr, P. H. S. (2016a). Struck: Structured Output Tracking with Kernels. *IEEE Trans. PAMI*, 38(10), 2096-2109.

Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M. M., Hicks, S. L. & Torr, P. H. S. (2016b). Struck: Structured Output Tracking with Kernels. *IEEE Trans. PAMI*, 38(10), 2096-2109.

He, A., Luo, C., Tian, X. & Zeng, W. A Twofold Siamese Network for Real-Time Object Tracking. *CVPR 2018*.

He, J., Mao, R., Shao, Z. & Zhu, F. (2020a). Incremental learning in online scenario. *CVPR*.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*.

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020b). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738.

He, L., Liang, J., Li, H. & Sun, Z. (2018a). Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. *CVPR*.

He, L., Sun, Z., Zhu, Y. & Wang, Y. (2018b). Recognizing partial biometric patterns. *CoRR*, abs/1810.07399.

He, L., Wang, Y., Liu, W., Zhao, H., Sun, Z. & Feng, J. (2019). Foreground-aware Pyramid Reconstruction for Alignment-free Occluded Person Re-identification. *ICCV*.

He, S., Luo, H., Wang, P., Wang, F., Li, H. & Jiang, W. (2021). Transreid: Transformer-based object re-identification. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013–15022.

Held, D., Thrun, S. & Savarese, S. Learning to Track at 100 FPS with Deep Regression Networks. *ECCV 2016*.

Henriques, J. F., Caseiro, R., Martins, P. & Batista, J. (2015). High-Speed Tracking with Kernelized Correlation Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 583-596. doi: 10.1109/TPAMI.2014.2345390.

Hermans, A., Beyer, L. & Leibe, B. (2017a). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hermans, A., Beyer, L. & Leibe, B. (2017b). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hirzer, M., Roth, P. M., Köstinger, M. & Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. *ECCV*.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Hong, C., Yu, J., Zhang, J., Jin, X. & Lee, K.-H. (2018). Multimodal face-pose estimation with multitask manifold deep learning. *IEEE transactions on industrial informatics*, 15(7), 3952–3961.

Hornakova, A., Henschel, R., Rosenhahn, B. & Swoboda, P. (2020). Lifted disjoint paths with application in multiple object tracking. *International conference on machine learning*, pp. 4364–4375.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S. & Chen, X. (2019a). Vrstc: Occlusion-free video person re-identification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7183–7192.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S. & Chen, X. (2019b, June). VRSTC: Occlusion-Free Video Person Re-Identification. *CVPR*.

Hou, R., Chang, H., Ma, B., Shan, S. & Chen, X. (2020). Temporal complementary learning for video person re-identification. *European Conference on Computer Vision*, pp. 388–405.

Hu, J., Shen, L. & Sun, G. (2018, June). Squeeze-and-Excitation Networks. *CVPR*, pp. 7132-7141. doi: 10.1109/CVPR.2018.00745.

Hu, J., Shen, L. & Sun, G. (2018a). Squeeze-and-excitation networks. *CVPR*, pp. 7132–7141.

Hu, J., Shen, L. & Sun, G. (2018b). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Huang, B., Chen, J., Xu, T., Wang, Y., Jiang, S., Wang, Y., Wang, L. & Li, J. (2021, October). SiamSTA: Spatio-Temporal Attention Based Siamese Tracker for Tracking UAVs. *ICCV*, pp. 1204-1212.

Huang, C., Loy, C. C. & Tang, X. (2016). Local similarity-aware deep feature embedding. *Advances in neural information processing systems*, 29.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017). Densely connected convolutional networks. *CVPR*.

Huang, H., Li, D., Zhang, Z., Chen, X. & Huang, K. (2018a). Adversarially occluded samples for person re-identification. *CVPR*.

Huang, H., Li, D., Zhang, Z., Chen, X. & Huang, K. (2018b). Adversarially occluded samples for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5098–5107.

Huang, L., Zhao, X. & Huang, K. (2019). Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hui, T.-W., Tang, X. & Loy, C. C. (2018). LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. *CVPR*, pp. 8981–8989.

Ilse, M., Tomczak, J. M. & Welling, M. (2018). Attention-based Deep Multiple Instance Learning. *arXiv preprint arXiv:1802.04712*.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, pp. 448–456.

Izutov, E. (2018). Fast and accurate person re-identification with rmnet. *arXiv preprint arXiv:1812.02465*.

Jacobs, D. W., Weinshall, D. & Gdalyahu, Y. (2000). Classification with non-metric distances: Image retrieval and class representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6), 583–600.

Ji, S., Xu, W., Yang, M. & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 221–231.

Jia, M., Cheng, X., Zhai, Y., Lu, S., Ma, S., Tian, Y. & Zhang, J. (2021). Matching on sets: Conquer occluded person re-identification without alignment. *AAAI*.

Jiang, B., Luo, R., Mao, J., Xiao, T. & Jiang, Y. (2018). Acquisition of localization confidence for accurate object detection. *Proceedings of the European conference on computer vision (ECCV)*, pp. 784–799.

Jin, Q., Han, Y., Wang, W., Tang, L., Li, J. & Deng, C. (2024). An Occlusion-Aware Tracker with Local-Global Features Modeling in UAV Videos. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Jin, X., Lan, C., Zeng, W., Wei, G. & Chen, Z. (2020). Semantics-aligned representation learning for person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11173–11180.

Jung, I., You, K., Noh, H., Cho, M. & Han, B. (2020). Real-time object tracking via meta-learning: Efficient model adaptation and one-shot channel pruning. *AAAI*, 34(07).

Kalal, Z., Mikolajczyk, K. & Matas, J. (2012). Tracking-Learning-Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7), 1409–1422. doi: 10.1109/TPAMI.2011.239.

Kalayeh, M., Basaran, E., Gökmen, M., Kamasak, M. E. & Shah, M. (2018). Human semantic parsing for person re-identification. *CVPR*.

Karanam, S., Li, Y. & Radke, R. J. (2015). Person re-identification with discriminatively trained viewpoint invariant dictionaries. *Proceedings of the IEEE international conference on computer vision*, pp. 4516–4524.

Kendall, A., Gal, Y. & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ayed, I. B. (2019). Constrained-CNN losses for weakly supervised segmentation. *Medical image analysis*, 54, 88–99.

Kim, S.-W. & Duin, R. P. (2011). Dissimilarity-based classifications in eigenspaces. *Iberoamerican Congress on Pattern Recognition*, pp. 425–432.

Kiran, M., Nguyen-Meidine, L. T., Sahay, R., Cruz, R. M. O. E., Blais-Morin, L.-A. & Granger, E. Generative Target Update for Adaptive Siamese Tracking. *ICPRAI 2022*, pp. 502–513.

Kiran, M., Tiwari, V., Morin, L.-A. B., Granger, E. et al. (2019a). On the interaction between deep detectors and Siamese trackers in video surveillance. *AVSS*, pp. 1–8.

Kiran, M., Tiwari, V., Morin, L.-A. B., Granger, E. et al. (2019b). On the interaction between deep detectors and siamese trackers in video surveillance. *AVSS 2019*, pp. 1–8.

Kiran, M., Bhuiyan, A., Blais-Morin, L.-A., Ayed, I. B., Granger, E. et al. (2021a). Flow guided mutual attention for person re-identification. *Image and Vision Computing*, 113, 104246.

Kiran, M., Praveen, R. G., Nguyen-Meidine, L. T., Belharbi, S., Blais-Morin, L.-A. & Granger, E. (2021b). Holistic guidance for occluded person re-identification. *arXiv preprint arXiv:2104.06524*.

Kiran, M., Sahay, R., Cruz, R. M. O. E., Blais-Morin, L.-A., Granger, E. et al. (2022). Dynamic template selection through change detection for adaptive Siamese tracking. *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13), 3521–3526.

Koch, G., Zemel, R., Salakhutdinov, R. et al. (2015). Siamese neural networks for one-shot image recognition. *ICML deep learning workshop*, 2(1).

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M. & Bischof, H. (2012). Large scale metric learning from equivalence constraints. *CVPR*.

Kour, G. & Saabne, R. (2014a). Fast classification of handwritten on-line Arabic characters. *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pp. 312–318. doi: 10.1109/SOCPAR.2014.7008025.

Kour, G. & Saabne, R. (2014b). Real-time segmentation of on-line handwritten arabic script. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 417–422.

Krause, J., Stark, M., Deng, J. & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561.

Kristan, e. (2021). The Ninth Visual Object Tracking VOT2021 Challenge Results. *ICCV Workshop 2021)*, pp. 2711-2738.

Kristan, M. & et al. The Visual Object Tracking VOT2017 Challenge Results. *ICCVW 2017*.

Kristan, M. & et al. (2018). The sixth Visual Object Tracking VOT2018 challenge results.

Krizhevsky, A. & Hinton, G. E. (2011). Using very deep autoencoders for content-based image retrieval. *ESANN*, 1, 2.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.

Lai, S., Chai, Z. & Wei, X. (2021). Transformer meets part model: Adaptive part division for person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4150–4157.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S. & Schindler, K. (2015). MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *arXiv:1504.01942 [cs]*. arXiv: 1504.01942.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.

LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P. et al. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261(276), 2.

Li, B., Yan, J., Wu, W., Zhu, Z. & Hu, X. High Performance Visual Tracking With Siamese Region Proposal Network. *CVPR 2018*.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J. & Yan, J. (2019a). Siamrpn++: Evolution of siamese visual tracking with very deep networks. *CVPR*.

Li, D., Chen, X., Zhang, Z. & Huang, K. (2017a). Learning deep context-aware features over body and latent parts for person re-identification. *CVPR*.

Li, H., Li, Y. & Porikli, F. (2016). DeepTrack: Learning Discriminative Feature Representations Online for Robust Visual Tracking. *IEEE Trans. IP*, 25(4), 1834-1848.

Li, H., Li, Y. & Porikli, F. DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking. *BMVC 2014*.

Li, J., Wang, J., Tian, Q., Gao, W. & Zhang, S. (2019b). Global-local temporal representations for video person re-identification. *CVPR*, pp. 3958–3967.

Li, J., Zhang, S. & Huang, T. (2019c). Multi-scale 3d convolution network for video based person re-identification. *AAAI*, 33, 8618–8625.

Li, J., Zhang, S. & Huang, T. (2020). Multi-scale temporal cues learning for video person re-identification. *IEEE Transactions on Image Processing*, 29, 4461–4473.

Li, W., Zhu, X. & Gong, S. (2018a). Harmonious attention network for person re-identification. *CVPR*.

Li, W., Zhao, R., Xiao, T. & Wang, X. (2014a). Deepreid: Deep filter pairing neural network for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159.

Li, W., Zhao, R., Xiao, T. & Wang, X. (2014b). Deepreid: Deep filter pairing neural network for person re-identification. *CVPR*.

Li, W., Zhu, X. & Gong, S. (2017b). Person re-identification by deep joint learning of multi-loss classification. *IJCAI*.

Li, Y., Liu, Y., Zhang, H., Zhao, C., Wei, Z. & Miao, D. (2024). Occlusion-Aware Transformer with Second-Order Attention for Person Re-Identification. *IEEE Transactions on Image Processing*.

Li, Y. & Zhang, X. (2019). SiamVGG: Visual Tracking using Deeper Siamese Networks.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y. & Wu, F. (2021a). Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer. *CVPR*.

Li, Z., Shao, H., Niu, L. & Xue, N. (2021b). Progressive learning algorithm for efficient person re-identification. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 16–23.

Li, Z., Bilodeau, G.-A. & Bouachi, W. (2018b). Multi-Branch Siamese Networks with Online Selection for Object Tracking. *arXiv preprint arXiv:1808.07349*.

Li Zhang, Yuan Li & Nevatia, R. (2008a, June). Global data association for multi-object tracking using network flows. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8. doi: 10.1109/CVPR.2008.4587584.

Li Zhang, Yuan Li & Nevatia, R. (2008b, June). Global data association for multi-object tracking using network flows. *CVPR*, pp. 1-8. doi: 10.1109/CVPR.2008.4587584.

Liao, S., Jain, A. & Li, S. (2012). Partial face recognition: Alignment-free approach. *PAMI*, 35(5), 1193–1205.

Liao, S., Hu, Y., Zhu, X. & Li, S. (2015). Person re-identification by local maximal occurrence representation and metric learning. *CVPR*.

Liao, S. & Li, S. Z. (2015). Efficient psd constrained asymmetric metric learning for person re-identification. *ICCV*.

Lin, T., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017, Oct). Focal Loss for Dense Object Detection. *ICCV*, pp. 2999-3007. doi: 10.1109/ICCV.2017.324.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014a). Microsoft COCO: Common Objects in Context. *ECCV*, pp. 740–755.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014b). Microsoft coco: Common objects in context. *ECCV*.

Lisanti, G., Masi, I., Bagdanov, A. D. & Del Bimbo, A. (2015). Person re-identification by iterative re-weighted sparse ranking. *CVPR*.

Liu, C.-T., Wu, C.-W., Wang, Y.-C. F. & Chien, S.-Y. (2019a). Spatially and Temporally Efficient Non-local Attention Network for Video-based Person Re-Identification. *BMVC*.

Liu, C., Gong, S., Loy, C. C. & Lin, X. (2012). Person re-identification: What features are important? *ECCV*.

Liu, H., Feng, J., Qi, M., Jiang, J. & Yan, S. (2017a). End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7), 3492–3506.

Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S. & Hu, J. (2018). Pose transferrable person re-identification. *CVPR*.

Liu, J., Shahroudy, A., Xu, D. & Wang, G. (2016). Spatio-temporal LSTM with trust gates for 3D human action recognition. *ECCV*.

Liu, Q., Chen, D., Chu, Q., Yuan, L., Liu, B., Zhang, L. & Yu, N. (2022). Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing*, 483, 333–347.

Liu, S., Liang, Y. & Gitter, A. (2019b). Loss-balanced task weighting to reduce negative transfer in multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 9977–9978.

Liu, S., Johns, E. & Davison, A. J. (2019c). End-to-end multi-task learning with attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1871–1880.

Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J. & Wang, X. (2017b). Hydraplus-net: Attentive deep features for pedestrian analysis. *ICCV*.

Liu, Y., Junjie, Y. & Ouyang, W. (2017c). Quality Aware Network for Set to Set Recognition. *CVPR*.

Lobo, J. L., Laña, I., Del Ser, J., Bilbao, M. N. & Kasabov, N. (2018). Evolving spiking neural networks for online learning over drifting data streams. *Neural Networks*, 108, 1–19.

Lopez-Paz, D. & Ranzato, M. (2017a). Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 6467–6476.

Lopez-Paz, D. & Ranzato, M. (2017b). Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 6467–6476.

Loy, C. C., Xiang, T. & Gong, S. (2009). Multi-camera activity correlation analysis. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1988–1995.

Luo, H., Gu, Y., Liao, X., Lai, S. & Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0.

Ma, C.-Y., Chen, M.-H., Kira, Z. & AlRegib, G. (2019). TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, 71, 76–87.

Ma, H., Li, X., Yuan, X. & Zhao, C. (2023). Two-phase self-supervised pretraining for object re-identification. *Knowledge-Based Systems*, 261, 110220.

Mahmoudi, N., Ahadi, S. M. & Rahmati, M. (2019). Multi-target tracking using CNN-based features: CNNMTT. *Multimedia Tools and Applications*, 78(6), 7077–7096.

Maninis, K.-K., Radosavovic, I. & Kokkinos, I. (2019). Attentive single-tasking of multiple tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1851–1860.

mAuth1. (2001). mTit1. *mJour1*, 1(1), 42-43.

mAuth2. (2002). mTit2. *mJour2*, 2(2), 42-43.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115–133.

McLaughlin, N., d. Rincon, J. M. & Miller, P. (2016a, June). Recurrent Convolutional Network for Video-Based Person Re-identification. *CVPR*, pp. 1325-1334. doi: 10.1109/CVPR.2016.148.

McLaughlin, N., d. Rincon, J. M. & Miller, P. (2016b, June). Recurrent Convolutional Network for Video-Based Person Re-identification. *CVPR*, pp. 1325-1334. doi: 10.1109/CVPR.2016.148.

Mekhazni, D., Bhuiyan, A., Ekladious, G. & Granger, E. (2020). Unsupervised domain adaptation in the dissimilarity space for person re-identification. *European Conference on Computer Vision*, pp. 159–174.

Miao, J., Wu, Y., Liu, P., Ding, Y. & Yang, Y. (2019). Pose-guided feature alignment for occluded person re-identification. *ICCV*.

Migneault, F. C., Granger, E. & Mokhayeri, F. (2018). Using Adaptive Trackers for Video Face Recognition from a Single Sample Per Person. *IPTA 2018*, pp. 1-6. doi: 10.1109/IPTA.2018.8608163.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S. & Schindler, K. (2016). MOT16: A Benchmark for Multi-Object Tracking. *arXiv:1603.00831 [cs]*. arXiv: 1603.00831.

Ming, Z., Zhu, M., Wang, X., Zhu, J., Cheng, J., Gao, C., Yang, Y. & Wei, X. (2022). Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing*, 119, 104394.

Ming, Z., Chazalon, J., Luqman, M. M., Visani, M. & Burie, J.-C. (2017). Simple triplet loss based on intra/inter-class metric learning for face verification. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1656–1664.

Mirmahboub, B., Kiani, H., Bhuiyan, A., Perina, A., Zhang, B., Del Bue, A. & Murino, V. (2016). Person re-identification using sparse representation with manifold constraints. *ICIP*.

Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S. & Singh, S. (2017). No fuss distance metric learning using proxies. *Proceedings of the IEEE international conference on computer vision*, pp. 360–368.

Mueller, M., Smith, N. & Ghanem, B. (2016). A benchmark and simulator for uav tracking. *European conference on computer vision*, pp. 445–461.

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S. & Ghanem, B. (2018). Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. *ECCV*.

Musgrave, K., Belongie, S. & Lim, S.-N. (2020). A metric learning reality check. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 681–699.

Nam, H. & Han, B. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. *CVPR 2016*.

Nam, H., Baek, M. & Han, B. (2016). Modeling and propagating cnns in a tree structure for visual tracking. *arXiv:1608.07242*.

Nanni, L., Rigo, A., Lumini, A. & Brahnam, S. (2020). Spectrogram classification using dissimilarity space. *Applied Sciences*, 10(12), 4176.

Nebehay, G. & Pflugfelder, R. Consensus-based matching and tracking of keypoints for object tracking. *WACV 2014*. doi: 10.1109/WACV.2014.6836013.

Nguyen-Meidine, L. T., Granger, E., Kiran, M. & Blais-Morin, L. A comparison of CNN-based face and head detectors for real-time video surveillance applications. *IPTA 2107*.

Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C. & He, Z. (2017, May). Spatially supervised recurrent convolutional neural networks for visual object tracking. *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-4. doi: 10.1109/ISCAS.2017.8050867.

Oliveira, L. E. Yandre MG Costa, Diego Bertolini, Alceu S. Britto, George DC Cavalcanti.

Ondrašovič, M. & Tarábek, P. (2021). Siamese visual object tracking: A survey. *IEEE Access*, 9, 110149–110172.

Overett, G. & Petersson, L. (2009, June). Fast features for time constrained object detection. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 23-30.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2), 100–115.

Pan, X., Luo, P., Shi, J. & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via ibn-net. *ECCV*.

Panda, R., Bhuiyan, A., Murino, V. & Roy-Chowdhury, A. K. (2017). Unsupervised adaptive re-identification in open world dynamic camera networks. *CVPR*.

Pang, B., Li, Y., Zhang, Y., Li, M. & Lu, C. (2020). Tubetk: Adopting tubes to track multi-object in a one-step training model. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6308–6318.

Park, E. & Berg, A. C. (2018). Meta-tracker: Fast and robust online adaptation for visual object trackers. *ECCV)*.

Park, H. & Ham, B. (2020). Relation network for person re-identification. *Proceedings of the AAAI conference on artificial intelligence*, 34(07), 11839–11847.

Pedagadi, S., Orwell, J., Velastin, S. & Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. *CVPR*.

Pękalska, E., Duin, R. P. & Paclík, P. (2006). Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2), 189–208.

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T. & Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. *CVPR*.

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.-G. & Xue, X. (2018). Pose-normalized image generation for person re-identification. *ECCV*.

Quan, R., Dong, X., Wu, Y., Zhu, L. & Yang, Y. (2019). Auto-ReID: Searching for a Part-aware ConvNet for Person Re-Identification.

Rabanser, S., Günnemann, S. & Lipton, Z. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. *NIPS*, 32.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G. & Lampert, C. H. (2017). icarl: Incremental classifier and representation learning. *CVPR*.

Redmon, J. & Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242*.

Redmon, J. & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *arXiv*.

Remigereau, F., Mekhazni, D., Abdoli, S., Cruz, R. M., Granger, E. et al. (2022a). Knowledge distillation for multi-target domain adaptation in real-time person re-identification. *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3853–3557.

Remigereau, F., Mekhazni, D., Abdoli, S., Cruz, R. M., Granger, E. et al. (2022b). Knowledge distillation for multi-target domain adaptation in real-time person re-identification. *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3853–3557.

Ren, S., He, K., Girshick, R. & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137-1149. doi: 10.1109/TPAMI.2016.2577031.

Ren, S., He, K., Girshick, R. & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 91–99.

Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W. & Yang, M.-H. (2018). Gated fusion network for single image dehazing. *CVPR*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C. (2016a). Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *ECCV workshop*.

Ristani, E. & Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. *CVPR*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C. (2016b). Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *ECCVWK*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C. (2016c). Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *ECCVWK*.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2015a). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3), 211-252.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015b). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R. & Hadsell, R. (2016). Progressive neural networks. *arXiv:1606.04671*.

Salti, S., Cavallaro, A. & Stefano, L. D. (2012). Adaptive Appearance Modeling for Video Tracking: Survey and Evaluation. *IEEE Trans. IP*, 21(10), 4334-4348.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520.

Saquib Sarfraz, M., Schumann, A., Eberle, A. & Stiefelhagen, R. (2018). A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *ICCV*.

Schroff, F., Kalenichenko, D. & Philbin, J. (2015a). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Schroff, F., Kalenichenko, D. & Philbin, J. (2015b). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Schulter, S., Vernaza, P., Choi, W. & Chandraker, M. (2017, July). Deep Network Flow for Multi-object Tracking. *CVPR*, pp. 2730-2739. doi: 10.1109/CVPR.2017.292.

Sener, O. & Koltun, V. (2018). Multi-task learning as multi-objective optimization. *arXiv preprint arXiv:1810.04650*.

Sethi, T. S. & Kantardzic, M. (2017). On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems with Applications*, 82, 77–99.

Shen, Q., Qiao, L., Guo, J., Li, P., Li, X., Li, B., Feng, W., Gan, W., Wu, W. & Ouyang, W. (2022, June). Unsupervised Learning of Accurate Siamese Tracking. *CVPR*, pp. 8101-8110.

Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W. & Li, S. Z. (2016a). Embedding deep metric for person re-identification: A study against large variations. *European conference on computer vision*, pp. 732–748.

Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W. & Li, S. Z. (2016b). Embedding deep metric for person re-identification: A study against large variations. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 732–748.

Shin, H., Lee, J. K., Kim, J. & Kim, J. (2017). Continual learning with deep generative replay. *arXiv:1705.08690*.

Shuai, B., Berneshawi, A. G., Modolo, D. & Tighe, J. (2020). Multi-object tracking with siamese track-rcnn. *arXiv preprint arXiv:2004.07786*.

Si, J., Zhang, H., Li, C., Kuen, J., Kong, X., Kot, A. C. & Wang, G. (2018, June). Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-identification. *CVPR*, pp. 5363-5372. doi: 10.1109/CVPR.2018.00562.

Simonyan, K. & Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, pp. 568–576.

Simonyan, K. & Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinha, A., Chen, Z., Badrinarayanan, V. & Rabinovich, A. (2018). Gradient adversarial training of neural networks. *arXiv preprint arXiv:1806.08028*.

Siong, L. Y., Mokri, S. S., Hussain, A., Ibrahim, N. & Mustafa, M. M. (2009, Aug). Motion detection using Lucas Kanade algorithm and application enhancement. *2009 International Conference on Electrical Engineering and Informatics*, 02, 537-542. doi: 10.1109/ICEEI.2009.5254757.

Smeulders, A. W. M., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A. & Shah, M. (2014). Visual Tracking: An Experimental Survey. *IEEE Trans PAMI*, 36(7), 1442-1468.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Somers, V., De Vleeschouwer, C. & Alahi, A. (2023a). Body part-based representation learning for occluded person Re-Identification. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1613–1623.

Somers, V., De Vleeschouwer, C. & Alahi, A. (2023b). Body part-based representation learning for occluded person re-identification. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1613–1623.

Son, J., Baek, M., Cho, M. & Han, B. (2017, July). Multi-object Tracking with Quadruplet Convolutional Neural Networks. *CVPR*, pp. 3786-3795. doi: 10.1109/CVPR.2017.403.

Song, C., Huang, Y., Ouyang, W. & Wang, L. (2018, June). Mask-Guided Contrastive Attention Model for Person Re-identification. *CVPR*, pp. 1179-1188. doi: 10.1109/CVPR.2018.00129.

Song, C., Huang, Y., Ouyang, W. & Wang, L. (2018a). Mask-guided contrastive attention model for person re-identification. *CVPR*.

Song, G., Leng, B., Liu, Y., Hetang, C. & Cai, S. (2018b). Region-based quality estimation network for large-scale person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R. W. & Yang, M.-H. (2018c). Vital: Visual tracking via adversarial learning. *CVPR*.

Sosnovik, I., Moskalev, A. & Smeulders, A. W. (2021). Scale equivariance improves siamese tracking. *WACV*, pp. 2765–2774.

Souza, V. L., Oliveira, A. L., Cruz, R. M. & Sabourin, R. (2019). On dissimilarity representation and transfer learning for offline handwritten signature verification. *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9.

Souza, V. L., Oliveira, A. L., Cruz, R. M. & Sabourin, R. (2020). A white-box analysis on the writer-independent dichotomy transformation applied to offline handwritten signature verification. *Expert Systems with Applications*, 154, 113397.

Souza, V. L., Oliveira, A. L., Cruz, R. M. & Sabourin, R. (2021). An investigation of feature selection and transfer learning for writer-independent offline handwritten signature verification. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7478–7485.

Srivastava, R. K., Greff, K. & Schmidhuber, J. (2015). Training very deep networks. *NIPS*, pp. 2377–2385.

Stollenga, M. F., Masci, J., Gomez, F. & Schmidhuber, J. (2014). Deep networks with internal selective attention through feedback connections. *NIPS*.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W. & Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. *ICCV*.

Subramaniam, A., Nambiar, A. & Mittal, A. (2019a). Co-segmentation inspired attention networks for video-based person re-identification. *CVPR*.

Subramaniam, A., Nambiar, A. & Mittal, A. (2019b, October). Co-Segmentation Inspired Attention Networks for Video-Based Person Re-Identification. *ICCV*.

Subramaniam, A., Nambiar, A. & Mittal, A. (2019c). Co-segmentation inspired attention networks for video-based person re-identification. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 562–572.

Suh, Y., Wang, J., Tang, S., Mei, T. & Lee, K. (2018a). Part-Aligned Bilinear Representations for Person Re-identification. *ECCV*.

Suh, Y., Wang, J., Tang, S., Mei, T. & Lee, K. M. (2018b). Part-Aligned Bilinear Representations for Person Re-Identification. *ECCV*.

Sun, Y., Zheng, L., Yang, Y., Tian, Q. & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). *ECCV*.

Sun, Y., Zheng, L., Deng, W. & Wang, S. (2017). Svdnet for pedestrian retrieval. *ICCV*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*.

Tan, H., Liu, X., Tian, S., Yin, B. & Li, X. (2020). MHSA-Net: Multi-Head Self-Attention Network for Occluded Person Re-Identification. *CoRR*, abs/2008.04015.

Tang, S., Andriluka, M., Andres, B. & Schiele, B. (2017). Multiple people tracking by lifted multicut and person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3539–3548.

Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G. & Yang, J. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129, 123–130.

Tao, R., Gavves, E. & Smeulders, A. W. (2016). Siamese instance search for tracking. *CVPR*.

Tay, C.-P., Roy, S. & Yap, K.-H. (2019). AANet: Attribute Attention Network for Person Re-Identifications. *CVPR*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Valero-Mas, J. J., Gallego, A. J. & Rico-Juan, J. R. (2024). An overview of ensemble and feature learning in few-shot image classification using siamese networks. *Multimedia Tools and Applications*, 83(7), 19929–19952.

Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A. & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. *CVPR*, pp. 2805–2813.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Varior, R. R., Shuai, B., Lu, J., Xu, D. & Wang, G. (2016a). A siamese long short-term memory architecture for human re-identification. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 135–153.

Varior, R., Haloi, M. & Wang, G. (2016b). Gated siamese convolutional neural network architecture for human re-identification. *ECCV*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veit, A. & Belongie, S. (2018). Convolutional networks with adaptive inference graphs. *ECCV*.

Vu, T., Osokin, A. & Laptev, I. Context-aware CNNs for person head detection. *ICCV 2015*.

Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wang, F., Luo, L. & Zhu, E. (2021a). Two-Stage Real-Time Multi-object Tracking with Candidate Selection. *International Conference on Multimedia Modeling*, pp. 49–61.

Wang, F., Xiang, X., Cheng, J. & Yuille, A. L. (2017). Normface: L2 hypersphere embedding for face verification. *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049.

Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E. & Sun, J. (2020a). High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. *CVPR*.

Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E. & Sun, J. (2020b). High-order information matters: Learning relation and topology for occluded person re-identification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6449–6458.

Wang, G., Luo, C., Sun, X., Xiong, Z. & Zeng, W. (2020c). Tracking by instance detection: A meta-learning approach. *CVPR*.

Wang, G., Yuan, Y., Chen, X., Li, J. & Zhou, X. (2018a). Learning discriminative features with multiple granularities for person re-identification. *Proceedings of the 26th ACM international conference on Multimedia*, pp. 274–282.

Wang, H., Shen, J., Liu, Y., Gao, Y. & Gavves, E. (2022a). Nformer: Robust person re-identification with neighbor transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7297–7307.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. & Wu, Y. (2014, June). Learning Fine-Grained Image Similarity with Deep Ranking. *CVPR*, pp. 1386-1393. doi: 10.1109/CVPR.2014.180.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. et al. (2020d). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10), 3349–3364.

Wang, N., Zhou, W., Wang, J. & Li, H. (2021b). Transformer meets tracker: Exploiting temporal context for robust visual tracking. *CVPR*, pp. 1571–1580.

Wang, T., Gong, S., Zhu, X. & Wang, S. (2014). Person re-identification by video ranking. *European conference on computer vision*, pp. 688–703.

Wang, X., O'Brien, M., Xiang, C., Xu, B. & Najjaran, H. Real-time visual tracking via robust Kernelized Correlation Filter. *ICRA 2017*.

Wang, X., O'Brien, M., Xiang, C., Xu, B. & Najjaran, H. Real-time visual tracking via robust Kernelized Correlation Filter. *ICRA 2017*.

Wang, X., Tang, J., Luo, B., Wang, Y., Tian, Y. & Wu, F. (2021c). Tracking by joint local and global search: A target-aware attention-based approach. *IEEE transactions on neural networks and learning systems*, 33(11), 6931–6945.

Wang, X., Han, X., Huang, W., Dong, D. & Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5022–5030.

Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B. & Weinberger, K. Q. (2018b). Resource aware person re-identification across multiple resolutions. *CVPR*.

Wang, Y., Chen, Z., Wu, F. & Wang, G. (2018c). Person re-identification with cascaded pairwise convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1470–1478.

Wang, Z., Wang, L., Wu, T., Li, T. & Wu, G. (2022b). Negative sample matters: A renaissance of metric learning for temporal grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3), 2613–2623.

Wang, Z., Zheng, L., Liu, Y., Li, Y. & Wang, S. (2020e). Towards real-time multi-object tracking. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 107–122.

Wei, L., Zhang, S., Yao, H., Gao, W. & Tian, Q. (2017). Glad: Global-local-alignment descriptor for pedestrian retrieval. *ACM*.

Wei, L., Zhang, S., Gao, W. & Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 79–88.

Wei, S.-E., Ramakrishna, V., Kanade, T. & Sheikh, Y. (2016). Convolutional pose machines. *CVPR*.

Wei, W., Yang, W., Zuo, E., Qian, Y. & Wang, L. (2022). Person re-identification based on deep learning—An overview. *Journal of Visual Communication and Image Representation*, 82, 103418.

Weinberger, K. Q. & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Weller-Fahy, D. J., Borghetti, B. J. & Sodemann, A. A. (2014). A survey of distance and similarity measures used within network intrusion anomaly detection. *IEEE Communications Surveys & Tutorials*, 17(1), 70–91.

Wen, Y., Zhang, K., Li, Z. & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. *European conference on computer vision*, pp. 499–515.

Wiewel, F. & Yang, B. (2021). Entropy-based Sample Selection for Online Continual Learning. *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1477–1481.

Wu, C.-Y., Manmatha, R., Smola, A. J. & Krahenbuhl, P. (2017a). Sampling matters in deep embedding learning. *Proceedings of the IEEE international conference on computer vision*, pp. 2840–2848.

Wu, C.-Y., Manmatha, R., Smola, A. J. & Krahenbuhl, P. (2017b). Sampling matters in deep embedding learning. *Proceedings of the IEEE international conference on computer vision*, pp. 2840–2848.

Wu, D., Zhang, K., Zheng, S.-J., Hao, Y.-T., Liu, F.-Q., Qin, X., Cheng, F., Zhao, Y., Liu, Q., Yuan, C.-A. et al. (2019a). Random Occlusion Recovery for Person Re-identification. *Journal of Imaging Science and Technology*, 63(3), 30405–1.

Wu, G., Zhu, X. & Gong, S. (2019b). Spatio-Temporal Associative Representation for Video Person Re-Identification. *BMVC*, pp. 278.

Wu, L., Wang, Y., Gao, J. & Li, X. (2018a). Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognition*, 73, 275–288.

Wu, Q. & Chan, A. B. (2021). Meta-Graph Adaptation for Visual Object Tracking. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6.

Wu, Y., Lim, J. & Yang, M.-H. Online Object Tracking: A Benchmark. *CVPR 2103*.

Wu, Y., Lim, J. & Yang, M.-H. Online Object Tracking: A Benchmark. *CVPR 2013*.

Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W. & Yang, Y. (2018b, June). Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning. *CVPR*.

Wu, Z., Li, Y. & Radke, R. J. (2015). Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *TPAMI*.

Xing, E., Jordan, M., Russell, S. J. & Ng, A. (2002). Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15.

Xiong, F., Gou, M., Camps, O. & Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. *ECCV*.

Xu, J., Cao, Y., Zhang, Z. & Hu, H. (2019). Spatial-temporal relation networks for multi-object tracking. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3988–3998.

Xu, J., Zhao, R., Zhu, F., Wang, H. & Ouyang, W. (2018). Attention-aware compositional network for person re-identification. *CVPR*.

Xu, T., Feng, Z.-H., Wu, X.-J. & Kittler, J. (2020). AFAT: adaptive failure-aware tracker for robust visual object tracking. *arXiv:2005.13708*.

Xuan, H., Stylianou, A., Liu, X. & Pless, R. (2020). Hard negative examples are hard, but useful. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 126–142.

Yan, C., Pang, G., Bai, X., Liu, C., Ning, X., Gu, L. & Zhou, J. (2021). Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia*, 24, 1665–1677.

Yan, J., Luo, L., Deng, C. & Huang, H. (2023). Adaptive hierarchical similarity metric learning with noisy labels. *IEEE Transactions on Image Processing*, 32, 1245–1256.

Yang, T. & Chan, A. B. (2018a). Learning Dynamic Memory Networks for Object Tracking. *ECCV*.

Yang, T. & Chan, A. B. (2018b). Learning dynamic memory networks for object tracking. *ECCV*, pp. 152–167.

Yao, Y., Wu, X., Zhang, L., Shan, S. & Zuo, W. (2018). Joint representation and truncated inference learning for correlation filter based tracking. *ECCV*.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6), 2872–2893.

Ye, Z., Hong, C., Zeng, Z. & Zhuang, W. (2022). Self-Supervised Person Re-Identification with Channel-Wise Transformer. *2022 IEEE International Conference on Big Data (Big Data)*, pp. 4210–4217.

Yi, D., Lei, Z., Liao, S. & Li, S. (2014). Deep metric learning for person re-identification. *ICPR*.

Yin, H., Weng, L., Li, Y., Xia, M., Hu, K., Lin, H. & Qian, M. (2023). Attention-guided siamese networks for change detection in high resolution remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 117, 103206.

Yoon, J., Yang, E., Lee, J. & Hwang, S. J. (2018). Lifelong Learning with Dynamically Expandable Networks. *ICLR*.

Yu, B., Liu, T., Gong, M., Ding, C. & Tao, D. (2018). Correcting the triplet selection bias for triplet loss. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 71–87.

Yu, F., Li, W., Li, Q., Liu, Y., Shi, X. & Yan, J. (2016). Poi: Multiple object tracking with high performance detection and appearance feature. *European Conference on Computer Vision*, pp. 36–42.

Yu, R., Du, D., LaLonde, R., Davila, D., Funk, C., Hoogs, A. & Clipp, B. (2022a). Cascade transformers for end-to-end person search. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7267–7276.

Yu, R., Du, D., LaLonde, R., Davila, D., Funk, C., Hoogs, A. & Clipp, B. (2022b). Cascade transformers for end-to-end person search. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7267–7276.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K. & Finn, C. (2020). Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*.

Yuan, Y., Chen, W., Yang, Y. & Wang, Z. (2020). In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 354–355.

Zenke, F., Poole, B. & Ganguli, S. (2017). Continual learning through synaptic intelligence. *ICML*.

Zhang, G., Zhang, Y., Zhang, T., Li, B. & Pu, S. (2023a). PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14133–14142.

Zhang, L., Xiang, T. & Gong, S. (2016). Learning a discriminative null space for person re-identification. *CVPR*.

Zhang, L., Gonzalez-Garcia, A., Weijer, J. v. d., Danelljan, M. & Khan, F. S. (2019a). Learning the model update for siamese trackers. *ICCV*.

Zhang, R., Li, J., Sun, H., Ge, Y., Luo, P., Wang, X. & Lin, L. (2019b). Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, 28(10), 4870–4882.

Zhang, R., Zhang, H., Ning, X., Huang, X., Wang, J. & Cui, W. (2023b). Global-aware siamese network for change detection on remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 199, 61–72.

Zhang, S., Wen, L., Bian, X., Lei, Z. & Li, S. (2018a). Occlusion-aware R-CNN: detecting pedestrians in a crowd. *ECCV*.

Zhang, T., Wei, L., Xie, L., Zhuang, Z., Zhang, Y., Li, B. & Tian, Q. (2021). Spatiotemporal transformer for video-based person re-identification. *arXiv preprint arXiv:2103.16469*.

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C. & Sun, J. (2017a). Alignedreid: Surpassing human-level performance in person re-identification. *CoRR*, abs/1711.08184.

Zhang, X., Zhou, X., Lin, M. & Sun, J. (2018b). Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856.

Zhang, X., Dong, H., Hu, Z., Lai, W.-S., Wang, F. & Yang, M.-H. (2018c). Gated Fusion Network for Joint Image Deblurring and Super-Resolution. *BMVC*.

Zhang, Y., Wang, C., Wang, X., Zeng, W. & Liu, W. (2020a). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*.

Zhang, Y., Chu, G., Li, P., Hu, X. & Wu, X. (2017b). Three-layer concept drifting detection in text data streams. *Neurocomputing*, 260, 393–403.

Zhang, Y., Wang, L., Qi, J., Wang, D., Feng, M. & Lu, H. (2018d). Structured siamese network for real-time visual tracking. *ECCV*.

Zhang, Z. & Peng, H. (2019). Deeper and wider siamese networks for real-time visual tracking. *CVPR*.

Zhang, Z., Peng, H., Fu, J., Li, B. & Hu, W. (2020b). Ocean: Object-aware anchor-free tracking. *arXiv:2006.10721*.

Zhang, Z., Lan, C., Zeng, W. & Chen, Z. (2020c). Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-based Person Re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10407–10416.

Zhang, Z., Lan, C., Zeng, W. & Chen, Z. (2020d). Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10407–10416.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X. & Tang, X. (2017a). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. *CVPR*.

Zhao, L., Li, X., Zhuang, Y. & Wang, J. (2017b). Deeply-learned part-aligned representations for person re-identification. *ICCV*.

Zhao, L., Li, X., Zhuang, Y. & Wang, J. (2017c). Deeply-Learned Part-Aligned Representations for Person Re-Identification. *ICCV*, pp. 3219-3228.

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H. & Hua, X.-S. (2017d). Spatio-temporal autoencoder for video anomaly detection. *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1933–1941.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. & Tian, Q. (2015a). Scalable person re-identification: A benchmark. *ICCV*.

Zheng, L., Huang, Y., Lu, H. & Yang, Y. (2019a). Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28(9).

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. & Tian, Q. (2015b). Scalable Person Re-identification: A Benchmark. *Computer Vision, IEEE International Conference on*.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S. & Tian, Q. (2016). *MARS: A Video Benchmark for Large-Scale Person Re-identification*. European Conference on Computer Vision.

Zheng, L., Huang, Y., Lu, H. & Yang, Y. (2017a). Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*.

Zheng, W.-S., Gong, S. & Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. *CVPR*.

Zheng, W.-S., Li, X., Xiang, T., Liao, S., Lai, J. & Gong, S. (2015c). Partial person re-identification. *ICCV*.

Zheng, W., Chen, Z., Lu, J. & Zhou, J. (2019b). Hardness-aware deep metric learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 72–81.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y. & Kautz, J. (2019c). Joint discriminative and generative learning for person re-identification. *CVPR*.

Zheng, Z., Zheng, L. & Yang, Y. (2017b). A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1), 1–20.

Zheng, Z., Zheng, L. & Yang, Y. (2017c). A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1), 1–20.

Zheng, Z., Zheng, L. & Yang, Y. (2017d). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *ICCV*.

Zheng, Z., Zheng, L. & Yang, Y. (2018). Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y. & Kautz, J. (2019d). Joint discriminative and generative learning for person re-identification. *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2138–2147.

Zhong, B., Bai, B., Li, J., Zhang, Y. & Fu, Y. (2018). Hierarchical tracking by reinforcement learning-based searching and coarse-to-fine verifying. *IEEE Transactions on Image Processing*, 28(5), 2331–2341.

Zhong, W., Jiang, L., Zhang, T., Ji, J. & Xiong, H. (2020a). A part-based attention network for person re-identification. *Multimedia Tools and Applications*, 79.

Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. (2020b). Random Erasing Data Augmentation. *AAAI*.

Zhong, Z., Zheng, L., Cao, D. & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. *CVPR*.

Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. (2020c). Random erasing data augmentation. *AAAI*, 34(07), 13001–13008.

Zhou, J., Wang, P. & Sun, H. (2020a). Discriminative and robust online learning for siamese visual tracking. *AAAI*, 34(07), 13017–13024.

Zhou, K., Yang, Y., Cavallaro, A. & Xiang, T. (2019). Omni-scale feature learning for person re-identification. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3702–3712.

Zhou, S., Wu, J., Zhang, F. & Sehdev, P. (2020b). Depth occlusion perception feature analysis for person re-identification. *Pattern Recognition Letters*, 138, 617–623.

Zhou, Z., Huang, Y., Wang, W., Wang, L. & Tan, T. (2017, July). See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-identification. *CVPR*, pp. 6776-6785. doi: 10.1109/CVPR.2017.717.

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W. & Yang, M.-H. (2018a). Online multi-object tracking with dual matching attention networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 366–382.

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W. & Yang, M.-H. (2018b). Online multi-object tracking with dual matching attention networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 366–382.

Zhu, X., Bhuiyan, A., Mekhalfi, M. L. & Murino, V. (2017). Exploiting Gaussian mixture importance for person re-identification. *AVSS*.

Zhu, Z., Wang, Q., Bo, L., Wu, W., Yan, J. & Hu, W. Distractor-aware Siamese Networks for Visual Object Tracking. *ECCV 2018*.

Zhuo, J., Chen, Z., Lai, J. & Wang, G. (2018a). Occluded person re-identification. *ICME*.

Zhuo, J., Lai, J. & Chen, P. (2019). A Novel Teacher-Student Learning Framework For Occluded Person Re-Identification. *CoRR*, abs/1907.03253.

Zhuo, J., Chen, Z., Lai, J. & Wang, G. (2018b). Occluded person re-identification. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6.

Zottesso, R. H., Costa, Y. M., Bertolini, D. & Oliveira, L. E. (2018). Bird species identification using spectrogram and dissimilarity approach. *Ecological Informatics*, 48, 187–197.