Modèle basé sur le Transformer pour une détection robuste du visage des nourrissons et des enfants hospitalisés en utilisant des images RVB et thermiques

par

Toufik BOURAS

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE EN CONCENTRATION PERSONNALISÉE
M. Sc. A.

MONTRÉAL, LE 02 MAI 2024

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

Mme. Rita Noumeir, directrice de mémoire
Département de génie électrique à l'École de technologie supérieure

M. Philippe Jouvet, codirecteur
Unité de soins intensifs pédiatriques, CHU Sainte justine

M. Simon Drouin, président du jury
Département de génie logiciel et TI

Mme. Catherine Laporte, examinatrice externe
Département de génie électrique

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 16 AVRIL 2024

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

Modèle basé sur le Transformer pour une détection robuste du visage des nourrissons et des enfants hospitalisés en utilisant des images RVB et thermiques

Toufik BOURAS

RÉSUMÉ

La localisation faciale dans les vidéos des patients en Unité de Soins Intensifs Pédiatriques (USIP) est une étape essentielle dans plusieurs applications de surveillance non invasive des patients basée sur la vidéo. Ces applications vont de l'évaluation de la douleur du patient à partir de l'expression faciale à l'estimation des rythmes cardiaque et respiratoire à partir des caractéristiques faciales. La précision de la localisation de visages des patients peut avoir un impact sur la qualité de lour surveillance. Dans le cadre général, les modèles de détection de visage basés sur les réseaux de neurones convolutifs, tels que RetinaFace, atteignent une grande précision. Cependant, leur précision diminue considérablement lorsqu'ils sont appliqués en USIP ou en Unité de Soins Intensifs Néonatals (USIN). Cette baisse peut être attribuée à l'environnement clinique difficile, notamment le visage du patient occulté, les conditions d'éclairage variables et les postures extrêmes des patients.

Pour remédier à cela, nous utilisons un modèle de détection basé sur les Transformers, DEtection TRansformer (Detr), préentraîné sur l'ensemble des données WiderFace, pour détecter les visages en USIP. Nous avons également utilisé les images thermiques pour améliorer la précision de la détection faciale.

Nos résultats montrent que le modèle Detr se généralise très bien aux données USIP par rapport à RetinaFace. De plus, nous avons mis au point une approche novatrice intégrant des images RVB et thermiques légèrement alignées, ce qui a considérablement amélioré la précision de détection pour les modèles Detr et RetinaFace. En exploitant conjointement les images thermiques et RVB, un modèle Detr préentraîné a surpassé RetinaFace de 15,3 %, atteignant une précision moyenne de 71,6 %. Enfin, nous présentons les résultats de l'ajustement fin des deux modèles sur un ensemble de 282 images de divers patients, de différents âges et postures en USIP. Le modèle Detr basé sur le Transformer démontre une meilleure capacité de généralisation que le modèle RetinaFace basé sur les CNN pour la détection des visages en USIP.

Mots-clés: Enfants, réseaux neuronaux convolutifs, clinique, détection de visage, hôpital, Unité de Soins Intensifs Pédiatriques (USIP), pédiatrie, Transformer de vision, thermique, propriétés des Transformers, généralisation des Transformers

Transformer-Based Model for Robust Face Detection of Hospitalized Infants and Children Using RGB and Thermal Images

Toufik BOURAS

ABSTRACT

Face localization in videos of patients in the Pediatric Intensive Care Unit (PICU) is an essential step in several applications of video-based non-invasive patient monitoring. These applications range from assessing the patient's pain from facial expression to estimating the heart and respiratory rate from facial features. The localization accuracy of the patients' faces can impact the quality of the patient monitoring application. Currently, Convolutional Neural Network (CNN) based face detection models, such as RetinaFace, achieve high accuracy in general settings. However, their accuracy substantially declines when applied in the PICU or in the Neonatal Intensive Care Unit (NICU). Such decline can be attributed to the challenging clinical setting. Particularly, occluded patient face, variable lighting conditions, and extreme patient pose. Addressing this, we use a transformer-based detection model DEtection TRansformer (Detr) pre-trained on the WiderFace dataset to detect faces in the PICU. Our results show that the Detr model compared to RetinaFace generalizes very well to the PICU data. Moreover, we unveiled a novel approach integrating weakly aligned RGB and thermal images, boosting detection accuracy for both Detr and Retinaface. Leveraging both thermal and RGB images, a pre-trained Detr outperformed RetinaFace by 15.3% reaching an Average Precision (AP) of 71.6%. Finally, we discuss the results of fine-tuning both models on 282 images of diverse patients of different ages and poses in the PICU. The transformer-based model Detr generalizes better than the CNN-based RetinaFace model in detecting the faces in the PICU.

Keywords: Children, convolutional neural networks, clinic, face detection, hospital, PICU, pediatrics, vision transformer, thermal, transformer properties, transformer generalization

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AP	Average Precision
AUC	Area Under the Curve
CHU	Center Hospitalier Universitaire
CNN	Convolutional Neural Network
COMFORT-B	Comfort Behavioral
DenseNet	Dense Convolutional Network
Detr	DEtection TRansformer
FAN	Face Attention Network
FC	Fully Connected
FLACC	Face Legs Activity Cry Consolability
FPN	Feature Pyramid Network
GPU	Graphical Processing Unit
IA	Intelligence Artificielle
ICU	Intensive Care Unit
IEEE	The Institute of Electrical and Electronics Engineers
IOU	Intersection over Union
mAP	Mean Average Precision
MS COCO	Microsoft Common Objects in Context
MTCNN	Multitask Cascade Convolutional Neural Network
NMS	Non-Max Suppression
PAN	Path Aggregation Network
PASCAL VOC	The PASCAL Visual Object Classes
GCS	Glasgow Coma Scale
PICU	Pediatric Intensive Care Unit
RVB	Rouge Vert Bleu

ResNets	Residual Networks
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RPN	Region Proposal Network
SPP	Spatial Pyramid Pooling
SPP-Net	Spatial Pyramid Pooling Network
SVM	Support Vector Machine
USIN	Unités de Soins Intensifs Néonatals
USIP	Unité de Soins Intensifs Pédiatriques
VGG	Visual Geometry Group
YOLO	You Only Look Once

INTRODUCTION

Dans le cadre de cette introduction, nous exposerons les raisons qui nous ont poussés à développer un modèle de détection de visage robuste pour l'Unité de Soins Intensifs Pédiatriques (USIP). Nous débuterons en définissant la notion de conscience, puis nous expliquerons comment celle-ci est évaluée en milieu clinique à l'aide d'échelles comportementales. Ensuite, nous mettrons en évidence le fait que ces échelles requièrent une interaction directe avec les patients pour obtenir une évaluation précise de leur état. Cependant, il existe la possibilité d'évaluer passivement l'état du patient sans nécessiter une interaction directe, car la conscience est liée à diverses fonctions cérébrales, ce qui signifie qu'un changement de conscience est susceptible d'entraîner des modifications dans différentes fonctions corporelles, notamment les mouvements oculaires et les expressions faciales. Par conséquent, nous pouvons concevoir un outil automatique d'évaluation de la conscience basé sur les variations de la région faciale. Cependant, pour parvenir à une évaluation précise du niveau de conscience, il est impératif de détecter correctement la région faciale. Enfin, nous expliquerons pourquoi il est essentiel de disposer d'un modèle robuste pour la détection de la région faciale et comment nous développons ce modèle.

États de conscience

La conscience peut être définie comme une combinaison de l'éveil et de la perception. L'éveil réfère au niveau de vigilance de la personne, tandis que la perception reflète la capacité de la personne à se reconnaître elle-même et à percevoir son environnement (Cavanna, Shah, Eddy, Williams & Rickards, 2011). Des changements dans l'éveil, la perception, ou les deux, peuvent entraîner différentes états de conscience. Dans la Figure 0.1, nous illustrons différents états de conscience : l'état végétatif se caractérise par une perception très limitée (contenu) mais un niveau élevé d'éveil (niveau). À l'opposé de l'état végétatif, dans l'état de crise partielle complexe, la perception est très élevée tandis que l'éveil est très faible. L'état de conscience peut être altéré par divers facteurs, notamment un dysfonctionnement cérébral tel que le coma, des

facteurs physiologiques comme le sommeil, ou la présence de substances affectant la fonction cérébrale. En Unité de Soins Intensifs Pédiatriques (USIP), les patients reçoivent différentes substances sédatives et anesthésiantes pour réguler leur état de conscience. Dans ce contexte, la surveillance étroite de la conscience du patient est essentielle pour éviter les effets indésirables d'une sédation excessive ou insuffisante (Johansson & Kokinsky, 2009).

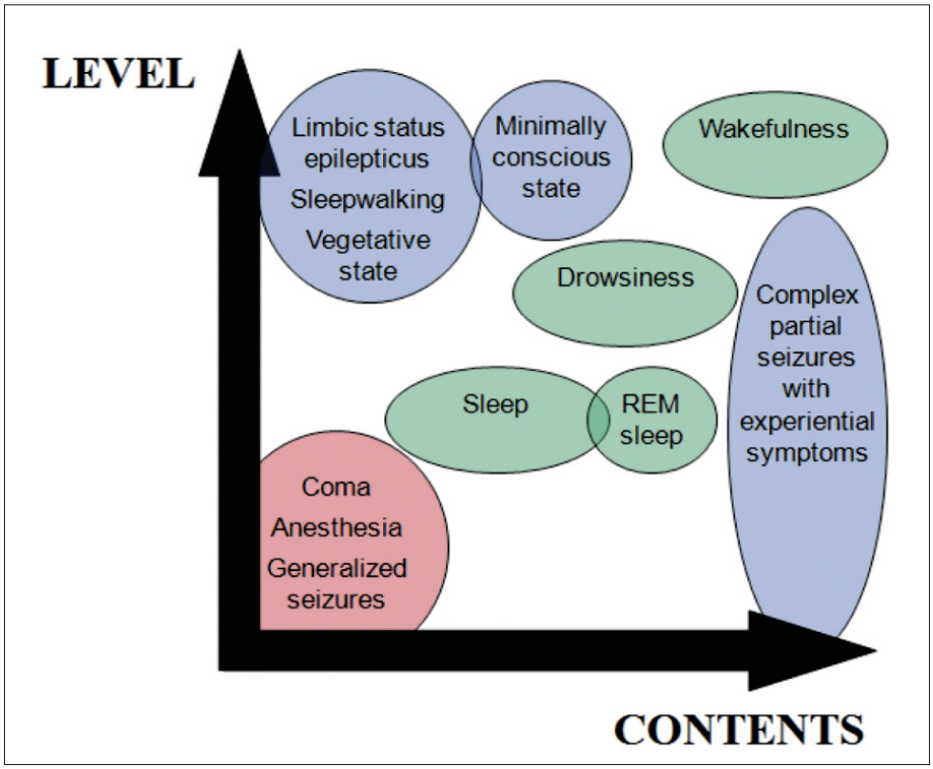

Figure 0.1 Conscience entre l'éveil (niveau) et la perception (contenu)
Tirée de Cavanna *et al.* (2011)

Échelles comportementales

Dans l'Unité de Soins Intensifs Pédiatriques (USIP), la surveillance continue de la conscience, de la douleur et d'autres signes vitaux des patients est essentielle pour détecter toute détérioration critique de leur santé. Certains de ces signes vitaux sont évalués à l'aide des appareils de mesure et de capteurs, tandis que d'autres sont évalués visuellement par des médecins et des infirmières formés sur différentes échelles d'évaluation des comportements. Les échelles comportementales

sont une composante essentielle de la surveillance des patients en USIP. Chaque échelle est spécifiquement conçue pour évaluer un aspect de l'état du patient. Voici quelques exemples :

- L'échelle FLACC (Sury & Bould, 2011) est utilisée pour mesurer le niveau de douleur chez les enfants en se basant sur leur expression faciale, leurs mouvements, leurs pleurs et d'autres comportements.
- L'échelle COMFORT-B (Ista, van Dijk, Tibboel & de Hoog, 2005) est une autre échelle d'évaluation de la douleur pour les enfants, qui inclut davantage d'indices comportementaux que FLACC. Elle est également utilisée pour évaluer le niveau de sédation chez les patients âgés de moins de 18 ans.
- L'échelle de Glasgow (GCS) (Sternbach, 2000) est employée pour évaluer le niveau de conscience des patients en fonction de leurs réponses motrices, verbales et de l'ouverture des yeux en réponse à des stimuli vocaux et sensoriels.

Ces échelles comportementales sont évaluées en termes de validité et de fiabilité. Une échelle est considérée valide si elle mesure effectivement ce qu'elle est censée mesurer, tandis qu'elle est considérée fiable si ses résultats sont cohérents entre différents évaluateurs et dans différents contextes. Lors de l'utilisation de ces échelles comportementales, le comportement de chaque patient est évalué en termes de rapidité, de qualité et de spontanéité (Sury & Bould, 2011). Si la spontanéité fait défaut, l'observateur peut appliquer un stimulus approprié, le cas échéant, en fonction de l'échelle utilisée.

Évaluation passive par l'analyse faciale

Les échelles comportementales telles que le GCS ont fait leurs preuves dans l'évaluation de la conscience des patients en milieu clinique, en se basant sur des indices comportementaux tels que la réponse motrice, la réponse verbale et l'ouverture des yeux. Cependant, ces échelles nécessitent des interactions directes avec le patient pour obtenir ces réponses. Par conséquent, le développement d'une évaluation automatique de la conscience, similaire aux échelles existantes,

peut être difficile pour plusieurs raisons. Tout d'abord, il peut être complexe de persuader le personnel médical de consentir à la collecte de données visuelles et auditives lors de leurs interactions avec les patients, même si ces données sont cruciales pour le développement d'un modèle d'Intelligence Artificielle (IA) visant à évaluer la conscience. Deuxièmement, un tel modèle ne fonctionnerait que lorsque des stimuli sensoriels similaires à l'interaction du personnel médical avec le patient sont présents. Heureusement, la conscience résulte de l'interaction de différentes parties du cerveau qui sont responsables de nombreuses fonctions corporelles. Par conséquent, un état de conscience altéré induit des modifications perceptibles dans les processus du corps humain. De nombreuses études ont démontré une relation significative entre les mouvements oculaires, y compris ceux associés au sommeil, ainsi que les réflexes oculaires, et le niveau de conscience des patients (Ting, Perez Velazquez & Cusimano, 2014; Wannez *et al.*, 2017; Cologan *et al.*, 2010). De plus, les rythmes respiratoires et cardiaques peuvent également être affectés en cas d'altération du niveau de conscience (Liuzzi *et al.*, 2023b,a). Par conséquent, il est envisageable, dans une certaine mesure, d'évaluer passivement le niveau de conscience des patients en combinant l'analyse de leurs mouvements oculaires, de leurs rythmes respiratoires, de leurs rythmes cardiaques et de leurs expressions faciales, qui seraient complètement absents dans l'état de coma extrême et très élevés dans l'état d'agitation. La possibilité d'évaluer passivement la conscience des patients ouvre la voie à la création d'un outil d'évaluation basé sur l'IA, qui utiliserait exclusivement le flux vidéo enregistré du patient. Le développement d'un tel outil implique généralement la détection précise d'une région d'intérêt à partir des images vidéo des patients, suivie de l'extraction des caractéristiques liées à la conscience. Dans cette recherche, nous avons choisi la région faciale comme région d'intérêt, car de nombreuses caractéristiques liées à la conscience peuvent être extraites de cette région. Par conséquent, notre première étape consiste à développer un modèle d'IA capable de détecter avec précision la région faciale à partir d'images montrant des patients dans différentes situations.

Description du problème

Au cours des années, la détection de visages a été l'objet de recherches approfondies, en particulier en raison de son importance pour une gamme variée d'applications, allant de l'identification faciale à des fins de sécurité à l'évaluation des expressions faciales pour détecter la douleur et d'autres réactions humaines. Le développement de la technologie de détection de visages est devenu possible grâce aux progrès des algorithmes avancés de détection des objets et à la disponibilité d'ensembles de données d'images mieux adaptés à l'environnement réel, caractérisé par la présence de personnes dans diverses situations et divers contextes. L'algorithme de (Viola & Jones, 2001) est l'un des premiers algorithmes de détection de visages à succès, montrant de bonnes performances à une vitesse relativement élevée. Cependant, il a ses limites, car il est plus performant pour détecter des visages en pose frontale et peu occultés. L'avènement des Réseaux Neuronaux Convolutifs (CNN) a permis de développer des modèles de détection de visages plus performants, capables de détecter des visages partiellement occlus, de différentes poses, de différentes tailles, et dans des conditions d'éclairage variées. Ces modèles de pointe basés sur des CNN, tels que RetinaFace (Deng, Guo, Ververas, Kotsia & Zafeiriou, 2020), ont une très grande précision lorsqu'ils sont entraînés et testés sur le même jeu de données comme le WiderFace (Yang, Luo, Loy & Tang, 2016). Cependant, leur performance diminue considérablement lorsqu'ils sont confrontés à des données provenant d'une distribution différente. Par exemple, lorsqu'ils sont testés sur des images de visages recueillies en clinique dans des Unités de Soins Intensifs Néonatals (USIN) ou en Unités de Soins Intensifs Pédiatriques (USIP), leur précision de détection est inférieure à celle initialement obtenue sur les données publiques de WiderFace (Yang *et al.*, 2016). Une solution potentielle à ce problème consiste à effectuer un ajustement fin du modèle préalablement entraîné sur des données cibles collectées dans un environnement clinique. Cependant, la collecte de données dans ce contexte présente des défis en raison de la petite taille de la population de patients et des difficultés associées à l'obtention du consentement pour la collecte d'images, ce qui limite la quantité de données de patients

disponibles. Dans cette mémoire, nous proposons une solution pour améliorer la précision de la détection de visages dans le contexte clinique en utilisant une technologie récemment développée : le Transformer de vision. Cette technologie présente certains avantages par rapport à la technologie de CNN utilisée précédemment. Les modèles basés sur le Transformer de vision ont récemment montré des performances supérieures dans différentes tâches liées à la vision (Han *et al.*, 2022), en partie grâce à leur architecture basée sur l'auto-Attention. Contrairement aux modèles CNN, les architectures courantes de Transformer de vision présentent moins de biais inductifs. Bien qu'ils nécessitent une grande quantité de données d'entraînement, ils sont très généralisables et requièrent moins de données pour l'ajustement fin lorsqu'ils sont adaptés à une tâche spécifique pertinente. Dans notre recherche, nous montrons comment le modèle de détection des objets basé sur le Transformer, Detr (Carion *et al.*, 2020), entraîné sur un ensemble de données public, le WiderFace, peut se généraliser efficacement à la détection de visages de patients en USIP, y compris dans des situations difficiles. Nous démontrons également l'avantage de combiner des données thermiques et RVB pour atteindre une meilleure précision de détection de visage des patients en USIP. Enfin, nous avons ajusté finement le DEtection TRansformer (Detr) et RetinaFace sur 282 images de patients en USIP. Nous avons réalisé les objectifs de cette recherche en deux étapes :

- Entraînement de Detr sur l'ensemble de données de WiderFace, qui comprend des images de visages, pour créer un modèle de base bien performant.
- Utilisation des images thermiques pour générer des images RVB plus ciblées des patients en USIP.

CHAPITRE 1

REVUE DE LITTÉRATURE

L'objectif de ce projet de recherche est de développer un modèle de localisation du visage du patient dans une image. Cet objectif peut être atteint en utilisant des modèles de détection de visage ou des objets. Habituellement, les architectures des modèles de détection de visage sont dérivées de réseaux de neurones génériques conçus pour la détection des objets. Par conséquent, il est essentiel de passer en revue les différents modèles de détection des objets à la pointe de la technologie afin de mieux comprendre les divers aspects liés à la localisation des objets.

1.1 Réseaux neuronaux convolutifs

Récemment, la plupart des modèles d'apprentissage profond réussis pour la détection des objets ont été développés en utilisant des couches convolutionnelles (Zou, Chen, Shi, Guo & Ye, 2023). Un modèle simple de détection des objets se construit en empilant plusieurs couches de convolution et de pooling, suivies de deux ou trois couches entièrement connectées. Dans les couches de convolution, un noyau avec des paramètres ajustables est utilisé pour extraire des caractéristiques à partir des données d'entrée. Pour illustrer, prenons l'exemple de la Figure 1.1, avec une image d'entrée en niveaux de gris représentée par une matrice 5x5. Un noyau de 3x3 peut être employé pour extraire des caractéristiques de l'image en effectuant une opération de convolution, qui consiste à multiplier élément par élément le noyau avec des sous-ensembles de l'image, également de taille 3x3 ;Les valeurs résultant de chaque opération de convolution sont ensuite additionnées pour produire une sortie unique. Le noyau est déplacé du coin de l'image, horizontalement puis verticalement, pour appliquer cette opération à tous les sous-ensembles de l'image. La matrice résultante de cette opération est appelée la carte des caractéristiques. Pendant l'entraînement, les paramètres du noyau de l'ensemble du réseau neuronal sont mis à jour afin de produire les meilleures caractéristiques pour la tâche de détection. Dans la couche de pooling, une opération similaire est appliquée à des sous-ensembles de la matrice d'entrée. Cependant, contrairement à la convolution, le pooling n'a aucun paramètre ajustable. Le pooling maximal est l'une des techniques de pooling les plus courantes, où la valeur maximale de chaque

sous-ensemble de l'entrée est sélectionnée. Après chaque opération de convolution, une fonction d'activation est appliquée à la carte des caractéristiques résultante. La sortie de cette fonction est appelée la carte des caractéristiques d'activation, et cette sortie est ensuite introduite dans une autre couche de convolution ou de pooling, en fonction de l'architecture du modèle.

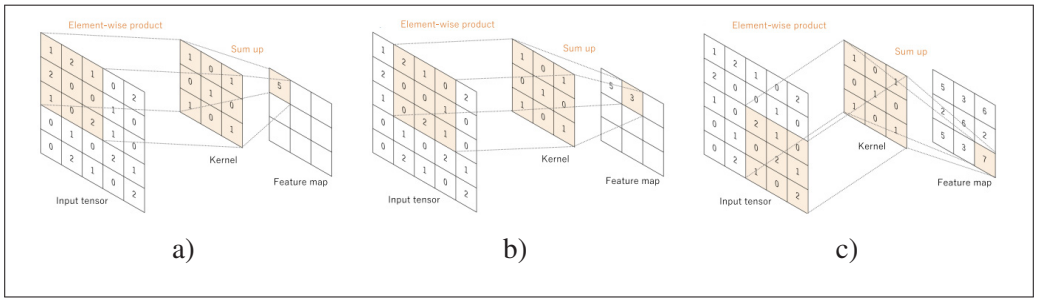

Figure 1.1 Exemple d'opération de convolution. a Le début de l'opération de convolution. b Déplacement du noyau d'un pas horizontalement. c La dernière étape de la convolution
Tirée de Yamashita *et al.* (2018)

Selon le résultat souhaité, les paramètres des couches de convolution et de pooling sont sélectionnés. Ces paramètres incluent la taille et le nombre de noyaux utilisés dans les couches de convolution, la taille du pooling, ainsi que le nombre de déplacements horizontaux et verticaux à effectuer à chaque étape de convolution ou de pooling.

Les couches empilées de CNN d'un modèle de détection d'objets sont appelées réseau neuronal d'extraction de caractéristiques. Dans la littérature, il porte différents noms : réseau convolutif, modèle de base, et modèle principal. Ce réseau neuronal est généralement adapté à partir de modèles bien connus développés pour la reconnaissance d'images : VGG (Simonyan & Zisserman, 2014), ResNet (He, Zhang, Ren & Sun, 2016), DenseNet (Huang, Liu, Van Der Maaten & Weinberger, 2017), GoogLeNet (Szegedy *et al.*, 2015), etc. De tels modèles, avec des architectures sophistiquées, préalablement entraînés sur des millions d'images, sont efficaces pour extraire des caractéristiques pertinentes. Cependant, lorsqu'ils sont utilisés sur des données dont la distribution diffère considérablement de celles sur lesquelles ils ont été entraînés, seules les couches initiales du réseau peuvent être utilisées efficacement. Les couches initiales sont spécialisées dans l'apprentissage de caractéristiques générales, telles que les bords et les taches

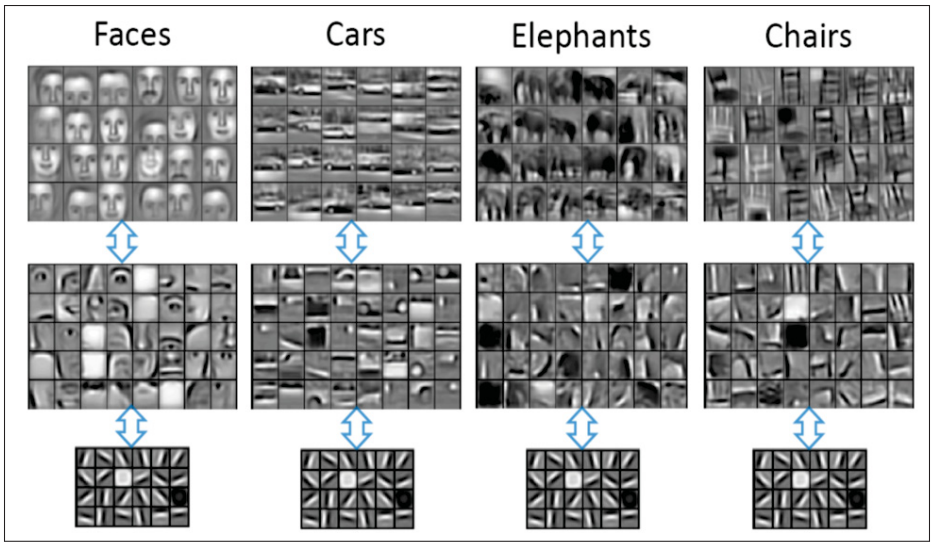

Figure 1.2 Illustration des caractéristiques détectées à différents niveaux du modèle d'apprentissage profond pour différents types de données d'entrée. Les couches initiales détectent des caractéristiques générales, mais les couches intermédiaires et les dernières couches détectent des caractéristiques plus pertinentes pour les données d'entrée Tirée de Gowtham (2018)

de couleur, applicables à la plupart des domaines. À mesure que l'on progresse des couches initiales aux couches finales, le réseau apprend des caractéristiques plus sophistiquées et spécifiques au jeu de données d'origine. La Figure 1.2 illustre les caractéristiques extraites par les couches initiales, intermédiaires et finales. Les couches initiales identifient des éléments tels que les bords de différentes orientations et d'autres caractéristiques générales, tandis que les couches intermédiaires se concentrent sur des caractéristiques plus pertinentes pour des parties spécifiques des données d'entrée, comme les traits du visage tels que les yeux et le nez. Les dernières couches sont responsables de la détection de caractéristiques globales de l'image d'entrée, telles que des visages ou des voitures.

1.2 Réseaux neuronaux Transformer

Les modèles basés sur le Transformer représentent la nouvelle génération de l'intelligence artificielle qui anime la vague actuelle d'applications réussies basées sur l'IA, telles que ChatGPT

(Ray, 2023). Contrairement aux modèles de vision basés sur les CNN (réseaux neuronaux convolutifs), où l'opération de convolution est appliquée à des sous-ensembles 2D de l'image d'entrée, les modèles de vision basés sur le Transformer utilisent le mécanisme d'auto-Attention sur des données d'entrée en 1D. Par conséquent, la matrice 2D de l'image est convertie en vecteurs 1D avant d'utiliser le mécanisme d'Attention pour extraire des caractéristiques à partir des vecteurs résultants. Dans le modèle Transformer de vision (ViT) (Dosovitskiy *et al.*, 2020), l'image d'entrée est divisée en N patchs de taille (P, P), et chaque patch est ensuite aplati en vecteurs de dimension D.

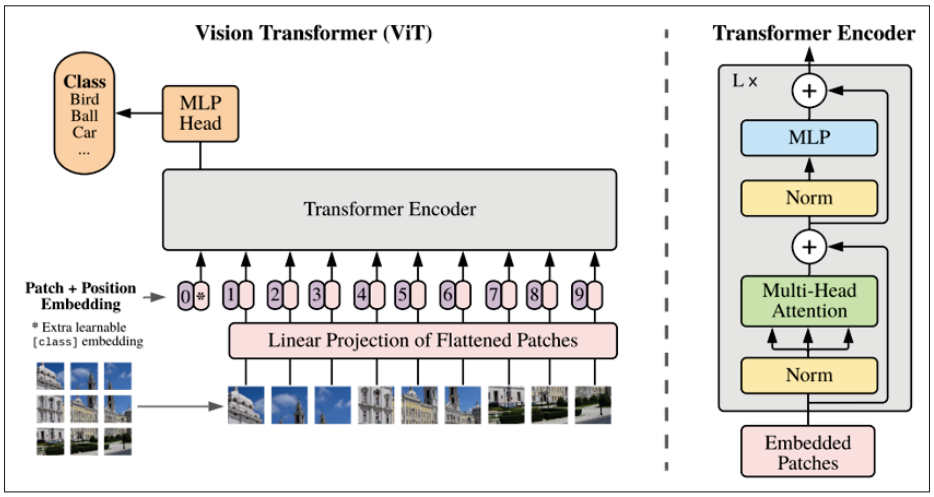

Figure 1.3 Architecture du modèle VIT
Tirée de Dosovitskiy *et al.* (2020)

Dans la Figure 1.3, on compte 9 patchs. Ensuite, une projection linéaire est appliquée à ces patchs pour obtenir leurs plongements. De plus, un plongement de classe supplémentaire, ajustable, est ajouté à la séquence des plongements des patchs. Enfin, un plongement positionnel ajustable est associé à chacun des plongements mentionnés précédemment. Ces plongements sont ensuite introduits dans l'encodeur du Transformer. À la sortie de l'encodeur, un perceptron multicouche (MLP) est appliqué pour accomplir la tâche de classification.

Le mécanisme d'Attention est au cœur de la technologie du modèle Transformer, comme illustré dans la Figure 1.4. Trois valeurs - Requête (Q), Clé (K) et Valeur (V) - sont estimées à partir de chaque patch d'entrée. Ces trois valeurs, Q, K et V, sont obtenues grâce à une projection

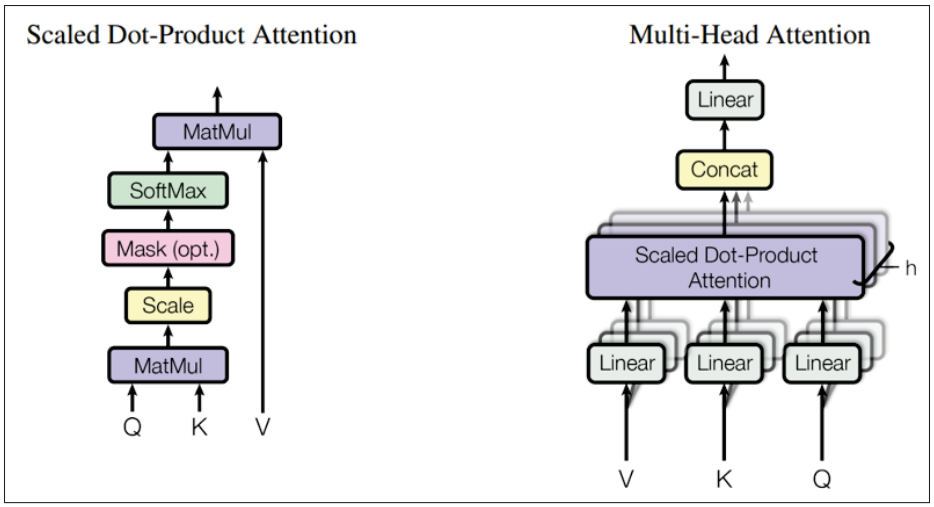

Figure 1.4 Le mécanisme d'Attention
Tirée de (Vaswani *et al.*, 2017)

linéaire des patchs, où les patchs sont respectivement multipliés par les matrices Wq, Wk et Wv. L'Attention par produit scalaire normalisé est une somme pondérée des valeurs, avec les poids calculés à partir des requêtes et des clés. L'équation d'Attention peut être formulée comme suit :

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}_\text{k}}}\right)\text{V}, \tag{1.1}$$

où d_k est la dimension de Q ainsi que de K.

Dans le produit scalaire de la fonction d'Attention, chaque requête de plongement d'entrée est multipliée à la fois par sa propre clé et par les clés des autres plongements d'entrée. Cela produit les poids de la fonction d'Attention. En conséquence, chaque sortie de la fonction d'Attention est liée à l'ensemble des plongements d'entrée. Cette relation est cruciale pour la capacité du modèle à obtenir de meilleures performances sur des entrées de grande taille.

Au lieu d'effectuer une seule opération d'Attention par un produit scalaire normalisé sur l'entrée, les auteurs Vaswani *et al.* (2017) ont suggéré de réaliser plusieurs opérations d'Attention parallèles

sur l'entrée en utilisant différentes projections linéaires Wq^i, Wk^i, Wv^i. Cette approche confère au modèle une plus grande flexibilité pour extraire des informations de l'entrée.

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})\text{W}^\text{O}, \tag{1.2}$$

où

$$\text{head}_\text{i} = \text{Attention}(\text{QW}_\text{i}^\text{Q}, \text{KW}_\text{i}^\text{K}, \text{VW}_\text{i}^\text{V}), \tag{1.3}$$

$\text{W}_\text{i}^\text{Q}$, $\text{W}_\text{i}^\text{K}$, $\text{W}_\text{i}^\text{V}$, et W^O sont les tenseurs de projection linéaire.

1.3 Multitâche dans la détection des objets

Un processus de détection de base implique deux tâches principales : la classification et la régression. La classification vise à distinguer l'objet détecté de l'arrière-plan et à identifier sa catégorie (par exemple, un visage), tandis que la régression est utilisée pour localiser précisément l'objet, soit directement, soit en utilisant les décalages de sa boîte d'ancrage(boîte prédéfinie fournie au modèle de detection). Bien que ces deux tâches puissent être abordées indépendamment en utilisant deux modèles distincts, cette configuration n'est pas optimale. En effet, les deux tâches exploitent des caractéristiques similaires extraites de l'image d'entrée, c'est pourquoi la plupart des modèles de détection des objets en apprentissage profond partagent des couches d'extraction de caractéristiques, mais disposent de couches distinctes pour la régression et la classification. Dans un modèle multitâche, la fonction de perte est une combinaison pondérée des différentes fonctions de perte associées à chaque tâche. Cette fonction est élaborée avec soin pour tenir compte des différences d'échelle entre la perte de classification et la perte de régression, de la vitesse de convergence de chaque tâche, ainsi que d'autres considérations liées au problème à résoudre. Lorsque les tâches sont similaires, les modèles multitâches peuvent souvent obtenir de meilleures performances, car ils apprennent une représentation des caractéristiques plus riche, améliorant ainsi la généralisation du modèle. De plus, l'utilisation d'un seul modèle pour plusieurs tâches au lieu de plusieurs modèles peut entraîner une réduction du nombre de paramètres, ce qui se traduit par des performances plus rapides. Toutefois, il est

important de noter que dans certains cas, les modèles multitâches peuvent sous-performer par rapport aux modèles séparés, phénomène connu sous le nom de "transfert négatif". Des détails supplémentaires sur l'apprentissage multitâche sont donnes par (Ruder, 2017).

1.4 Architectures de modèles de détection

Les architectures de détection des objets ont connu des évolutions au fil des années visant à améliorer à la fois la précision et la vitesse de détection. Dans cette section, nous présenterons les deux modèles de détection que nous avons utilisés dans notre étude. Le premier modèle, basé sur les CNN, est appelé RetinaFace, il partage des composants similaires avec un modèle classique de détection des objets. Le second modèle, nommé Detr, est un modèle général de détection des objets basé sur le Transformer.

1.4.1 Architecture du modèle RetinaFace

RetinaFace est l'un des meilleurs modèles de détection de visage, reconnu pour sa précision et sa vitesse compétitives. Ce modèle intègre de nombreuses avancées récentes dans le domaine de la détection des objets et de visages. Comme illustré dans la Figure 1.5, RetinaFace se compose de quatre composants principaux : modèle de base ResNet, un réseau de pyramide de caractéristiques, un module de contexte, et une fonction de perte multitâche. La connexion résiduelle du modèle de base Resnet permet d'utiliser des réseaux très profonds tout en limitant le problème perte de gradients associ à de tels modèles profonds. Ce backbone est pré-entraîné sur la classification à l'aide du vaste ensemble de données ImageNet, ce qui en fait un excellent extracteur de caractéristiques. La profondeur du backbone Resnet augmente le champ réceptif des caractéristiques en sortie, d'où une meilleure performance dans les grandes images. Le réseau de pyramide de caractéristiques agrège les caractéristiques extraites des couches profondes du backbone, ce qui permet de détecter des objets de différentes tailles en utilisant des caractéristiques de différents champs réceptifs. Le modèle utilise une tête de fonction de perte multitâche pour l'entraînement, comprenant une fonction perte de classification de visage, une fonction de perte de régression de boîte, une fonction de perte de régression de repères faciaux et une fonction

de perte de régression de visage dense. Il est entraîné sur l'ensemble de données WiderFace avec des techniques d'augmentation de données de base, telles que le recadrage aléatoire, le retournement horizontal et la distorsion de couleur photométrique.

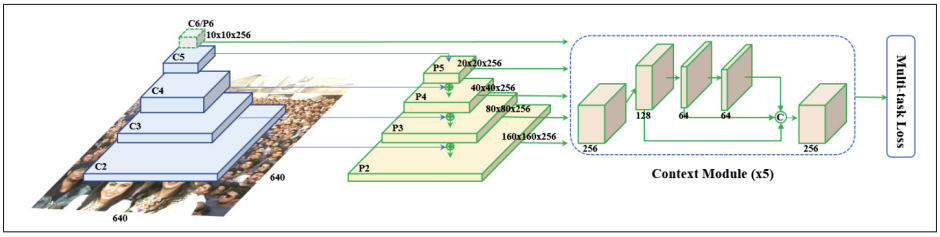

Figure 1.5 Architecture du modèle RetinaFace
Tirée de Deng *et al.* (2020)

1.5 Architecture du modèle Detr

Le modèle VIT a rencontré un succès initial dans la tâche de classification d'images, ouvrant ainsi la voie à l'exploration de nouvelles applications pour les modèles basés sur le Transformer dans le domaine de la vision par ordinateur. La détection des objets est l'une de ces tâches de vision importantes qui a connu un développement impressionnant depuis l'invention des CNN et qui est en cours de développement avec l'introduction de Transformer.

Detr est l'un des premiers modèles de détection basés sur le Transformer qui a montré une performance compétitive par rapport aux modèles de détection CNN. Ce modèle est hybride, ce qui signifie qu'il contient une combinaison de couches CNN et de Transformer.

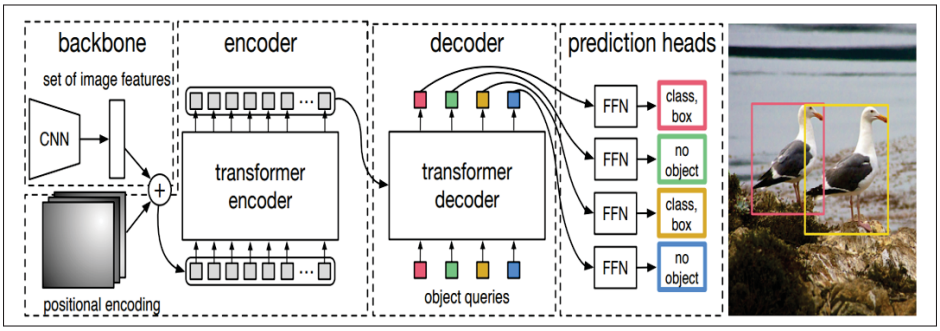

Figure 1.6 Architecture du modèle Detr
Tirée de Carion *et al.* (2020)

Comme illustré dans la Figure 1.6, les images d'entrée sont introduites dans le modèle CNN. Les caractéristiques de sortie du CNN sont aplaties en 1D avant d'être transmises au Transformer encodeur décodeur. Comme pour le modèle VIT, un codage positionnel est combiné aux caractéristiques d'entrée de l'encodeur. Les caractéristiques de sortie de l'encodeur sont combinées aux requêtes d'objet qui sont également un encodage positionnel. Le nombre N de sorties de caractéristiques du décodeur est le même que le nombre de requêtes d'objet. Ces caractéristiques sont ensuite utilisées séparément pour prédire la boîte englobante et la classe des objets présents dans l'image à différents emplacements.

CHAPITRE 2

L'APPROCHE ET L'ORGANISATION DU DOCUMENT

Le reste de ce document est structuré comme suit :

Chapitre 3 : l'article soumis pour publication dans le IEEE JOURNAL OF TRANSLATIONAL ENGINEERING IN HEALTH AND MEDICINE. Cet article examine les avantages de l'utilisation du modèle Transformer et des images thermiques pour améliorer la précision de la détection des visages en environnement de soins intensifs pédiatriques. Ce travail s'inscrit dans le cadre d'un effort plus vaste visant à développer un système de soutien à la décision clinique pour évaluer les signes vitaux des patients à l'aide de la caméra au chevet des patients. Pour atteindre les objectifs visés, l'article décrit d'abord l'entraînement du modèle de détection d'objets Detr sur l'ensemble de données WiderFace. Ensuite, il décrit comment nous avons utilisé les images thermiques pour améliorer la précision de la détection des visages.

Chapitre 4 : Discussion des implications de nos résultats sur la détection de visages en unités de soins intensifs pédiatriques (USIP) et au-delà de l'USIP.

Chapitre 5 : Dans la conclusion, nous avons inclus des recommandations pour poursuivre le développement de notre travail dans le contexte de l'environnement clinique.

CHAPITRE 3

TRANSFORMER-BASED MODEL FOR ROBUST FACE DETECTION OF HOSPITALIZED INFANTS AND CHILDREN USING RGB AND THERMAL IMAGES

Toufik Bouras[1] , Philippe Jouvet[2] , Rita Noumeir[1,2]

[1] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Groupe de recherche clinique en soins intensifs pediatriques, CHU Sainte-Justine,
3175 Chemin de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Paper submitted for publication in the *IEEE Journal of Translational Engineering in Health and Medicine*, October 2023

3.1 Abstract

Background : Face localization in videos of patients in the Pediatric Intensive Care Unit (PICU) is an essential step in several applications of video-based non-invasive patient monitoring. These applications range from assessing the patient's pain from facial expression to estimating the heart and respiratory rate from facial features. The localization accuracy of the patients' faces can impact the quality of the patient monitoring application. Currently, Convolutional Neural Network (CNN) based face detection models, such as RetinaFace, achieve high accuracy in general settings. However, their accuracy substantially declines when applied in the PICU or in the Neonatal Intensive Care Unit (NICU). Such decline can be attributed to the challenging clinical setting. Particularly, occluded patient face, variable lighting conditions, and extreme patient pose. Methods : Addressing this, we use a transformer-based detection model DEtection TRansformer (Detr) pre-trained on the WiderFace dataset to detect faces in the PICU. Results : Our results show that the Detr model compared to RetinaFace generalizes very well to the PICU data. Moreover, we unveiled a novel approach integrating weakly aligned RGB and thermal images, boosting detection accuracy for both Detr and Retinaface. Leveraging both thermal and RGB images, a pre-trained Detr outperformed RetinaFace by 15.3% reaching an Average Precision (AP) of 71.6%. Finally, we discuss the results of fine-tuning both models on 282 images of diverse patients of different ages and poses in the PICU. Conclusion : The transformer-based

model Detr generalizes better than the CNN-based RetinaFace model in detecting the faces in the PICU. Clinical and Translational Impact Statement : The generalization capability of the transformer model Detr makes it a better face detection model for clinical-related applications where the available clinical data for fine-tuning is limited.

3.2 Introduction

Continuous monitoring of critically ill children in the Pediatric Intensive Care Unit (PICU) is imperative to provide them with uninterrupted care and to promptly administer appropriate interventions should their health deteriorate. Conventionally, vital signs such as pulse rate and respiratory rate are monitored with wired sensors. Other signals such as discomfort and pain are monitored by doctors and nurses. Recently research has been done to develop contactless monitoring systems, amongst which systems that are based on video cameras. In these systems, usually, the Region of Interest (ROI) is selected in the image or in the frame of the video stream before further processing. The face as ROI is used in several studies to estimate, heart rate (Yang, He, Sadanand, Yusuf & Bolic, 2022), respiratory rate (Yang *et al.*, 2022), pain (Zamzmi *et al.*, 2016; Werner *et al.*, 2016), and discomfort (Sun *et al.*, 2019; Li, Pourtaherian, Van Onzenoort, a Ten & de With, 2020). Deploying the suggested applications of face detection within the unconstrained clinical setting presents some challenges. These challenges stem from factors such as face occlusions, varying lighting conditions, diverse poses, and the existence of objects resembling faces within the PICU room. Consequently, developing a more reliable face detection model is necessary for a broader adoption of such applications. The current state-of-the-art face detection models are CNN-based models. These models may not perform well when there is a substantial distribution difference between the source pre-training data and the target fine-tuning data. For instance, in (Dosso, Kyrollos, Greenwood, Harrold & Green, 2022), five pre-trained CNN-based face detection models showed a limited accuracy in detecting faces on images of Neonates in the Neonatal Intensive Care Unit (NICU). Consequently, collecting a considerable amount of data is necessary to successfully fine-tune the CNN-based face detection model on the target data. However, data collection in the clinical setting is challenging owing to factors

such as the limited patient pool, the difficulty in securing patient or guardian consent, and the intricacies involved in capturing data during medical procedures. Therefore, in this work, we propose to leverage the generalization capabilities of a transformer detection model to achieve better face detection accuracy than the CNN-based model RetinaFace. Specifically, we trained a transformer-based model on the WiderFace dataset (Yang *et al.*, 2016), then evaluated its performance on our PICU data. Moreover, we improved the obtained results by combining the information from thermal and RGB images. We further enhance the detection accuracy by fine-tuning both models RetinaFace and Detr on our PICU data collected at CHU Sainte-Justine Hospital (CHUSJ).

Automatic face detection is part of a broader endeavor to develop a clinical decision system (CDSS) at CHUJS. The facial features are essential in the development of an automated, video-based patient state assessment. As an integral part of the CDSS project, the face detection model will be instrumental in the automatic evaluation of various aspects of the patient's conditions including consciousness, pain, and discomfort levels. Upon detecting the facial region, an appropriate model will be developed to extract facial expression and eye movement features, as well as to infer small head movement features from a sequence of video frames of the patient's face. Subsequently, the extracted features will be used to evaluate patient conditions.

We review face detection methods used in different applications of patient monitoring of children and adults. We include studies of adult participants because these studies have comparable characteristics to the current study in terms of the clinical environment and the selection of the face region for clinical applications. For improved detection accuracy, some of the methods combine different modalities, thermal, RGB, depth, and others.

In (Zamzmi *et al.*, 2016), as a part of a multimodal approach to infant pain assessment, the Viola-Jones algorithm (Viola & Jones, 2001) which is one of the early successful face detection algorithms is used to detect the faces of hospitalized infants. The algorithm was able to detect the frontal and near frontal faces but failed to detect faces with extreme pose and occlusions. An

improved version of the Viola-Jones algorithm (Lienhart, Kuranov & Pisarevsky, 2003) was used by Werner *et al.* (2016) to detect the faces of adult participants from the RGB video frames. The participants were seated directly in front of the RGB and depth cameras. From each frame of the captured color and depth videos, head pose combined with facial expression features to estimate the pain of the participant over a period of the video registration. The quality of the face detection model was not discussed. However, the experiment was performed in a constrained environment, thus the faces could be easily detected.

Sun *et al.* (2019) used facial expression is used to detect discomfort from 55 videos of 24 infants. The infants were recorded during clinical procedures when they were experiencing distress moments. Only parts of the videos where the patients were in supine pose were selected, the recordings were conducted under varying lighting conditions. From each frame of the recording, the face was detected along the facial landmarks using an ensemble of regression trees (Kazemi & Sullivan, 2014). In this study, they used only the frames where facial landmarks were detected. An interesting approach for real-time infant discomfort monitoring was proposed by (Li *et al.*, 2020). They adapted the Faster-RCNN model, which is a general object detection model, to classify the discomfort of the infants and implicitly detect their faces. They used 16,165 infant images collected from the internet to train the model. Subsequently, the model was tested on ten two-minute videos of infants in the clinical environment and fifty videos from YouTube. They also trained their model on Near Infrared (NIR) images for detection at night. The proposed model showed a good discomfort detection performance of 89.9%, at an Average Precision (AP) corresponding to an Intersection over Union (IoU) threshold of 0.5 (AP@50). However, from the examples shown in the paper, the patients' faces were close to the camera with no or little occlusion.

Yang *et al.* (2022), proposed a system of two cameras, RGB and thermal, to monitor vital signs from faces covered with masks. RetinaFace was used to detect the facial region and the facial landmarks of the participants. After that, they aligned the thermal and the RGB images. Lastly, the forehead and the nostril regions were extracted to estimate Heart Rate (HR) from the RGB, respiration rate (RP), and body temperature (BT) from both the thermal and RGB. The accuracy

of the detection was not discussed quantitatively, but RetinaFace was qualitatively compared to other models. They demonstrated that RetinaFace outperformed the other models on masked face detection and landmark localization.

Kyrollos, Greenwood, Harrold & Green (2022) employed pressure images of neonates, captured via pressure-sensitive mats positioned beneath each patient, to refine a pre-trained head localization model. Given the challenge of directly labeling pressure images, they annotated the RGB images then using the geometric transformations they estimated the equivalent annotations for the pressure images. After training on 679 pressure images of five neonate patients, they achieved 54% at an AP@50. Gleichauf *et al.* (2023), proposed a method to fuse thermal with RGB data in order to train a robust neonatal facial region detector. Three cameras were placed near the neonate to capture thermal, RGB, and 3D Time of Flight (ToF) data. After calibration, they estimated the projection matrices between the three modalities. Subsequently, they fused thermal and RGB images. They trained RetinaFace (Deng *et al.*, 2020) and YOLOv3 (Redmon & Farhadi, 2018) on the three data modalities, RGB, thermal, and fused data. For the head detection task in the fused data, RetinaFace and YOLOv3 reached an AP of 99.58% and 94.55% respectively. The IoU threshold was not mentioned in their paper, but we assume they used the common threshold of 0.5. The achieved results are promising, although it is important to note that the patient cohort is notably small, consisting of only five individuals. Additionally, the video registration is performed in close range.

In (Dosso *et al.*, 2022), the complexity problem of the NICU environment is extensively analyzed for face detection applications. This study showed how the different occlusions, lighting conditions, and camera viewpoints impacted the face detection AP of five off-the-shelf models. Based on the results, RetinaFace and YOLOFace (Qi, Tan, Yao & Liu, 2022) were selected for further fine-tuning on the NICU data. Consequently, the AP@50 of RetinaFace and YOLOv5Face improved from 52.12% to 79.73% and from 48.85% to 87.98% respectively on the hard subset of their NICU data. These findings highlight the significance of fine-tuning face detection models specifically for clinical environments, instead of using them directly. While the obtained results are promising, it's important to note that their data consists of very

young demographic patients. The images of these patients were predominantly captured from short distances. Consequently, the faces in the images appeared relatively large. Moreover, it is worth mentioning that the dataset lacked the diversity of objects typically found in an ICU room.

Compared to the previous studies, our work is more aligned with the real-world PICU scenario both in terms of the data and the approach. We used images of 282 patients of different ages in the PICU. These patients have different health conditions. Therefore, different medical equipment such as breathing tubes, patches, electrodes, and heart monitors are used to treat, monitor, and provide comfort for these patients. One or more of these equipments cover parts of the patient's body, face, or both. Our approach consists of leveraging RGB and thermal images for face detection. In the thermal images, the patient region is less impacted by occlusion. Subsequently, the inclusion of the thermal image improves the model's accuracy and robustness. Moreover, we train a transformer-based model on the WiderFace before fine-tuning it on the PICU data. Transformer-based models are more robust to occlusion than CNN-based models (Naseer *et al.*, 2021). Finally, in this study, we present these two contributions :

1. We show that in contrast to RetinaFace, Detr model generalizes very well to the PICU data after being trained on the WiderFace. This result shows the potential of transformer-based models for applications in clinical settings beyond face detection. This generalization capability can be very useful when data is scarce which is a common problem in clinical applications.
2. We propose a method to leverage weakly aligned RGB and thermal images to improve detection accuracy.

3.2.1 LITERATURE REVIEW

3.2.1.1 OBJECT DETECTION

This section offers an overview of the development of deep learning-based object detection models, emphasizing their architectural advancements. It is important to highlight that early deep learning-based object detection models necessitated a multi-step training procedure. Additionally, these detection models have significantly benefited from multitask learning. Typically, the loss functions of these models are defined as a combination of task-specific losses, encompassing both classification and regression losses.

It has been demonstrated that face detection is fundamentally a special case of object detection where there is only one class to be detected, which is the face itself (Zhu, Cai, Zhang, Wang & Xiong, 2020). Modern object detection models can be categorized into two groups : one-stage detectors and two-stage detectors. The latter performs the detection in two steps : the region proposal step, and then the detection step.

One of the early two-stage deep learning models is Regions with CNN features (RCNN) (Girshick, Donahue, Darrell & Malik, 2014). In this model, from the input image, 2000 regions are proposed using the Selective Search Algorithm (SSA) (Uijlings, Van De Sande, Gevers & Smeulders, 2013). Then each region is warped to a fixed size. A CNN network is then applied to the resulting regions to extract the relevant features. These features are then classified using binary Support Vector Machines (SVM) (Hearst, Dumais, Osuna, Platt & Scholkopf, 1998). Based on the intersection over union with the ground truth some regions are selected as the right prediction. A bounding box regression is used to improve the detected region accuracy. RCNN is slow, hence, instead of extracting the features independently for each of the 2000 regions, SPPnets (He, Zhang, Ren & Sun, 2015) and Fast RCNN (Girshick, 2015) directly extract the features using a CNN. After that, SSA is applied to the resulting features. Moreover, in both models instead of region warping, pooling is applied to the extracted features in order to obtain a fixed-length representation to use for the classification layer. Faster RCNN (Ren, He, Girshick & Sun, 2015)

is developed to include region proposal as a part of the model architecture rather than a separate algorithm. In this model, the features resulting from the CNN are fed into two branches. First is the region proposal branch which contains a few layers to classify objects from the background and regress prior anchor boxes of different shapes and sizes. Based on a low IoU threshold, a soft selection of the predicted boxes is performed. The selected boxes are then subsequently utilized as regions for the detection branch. Faster RCNN has been among the best-performing detection models in both speed and accuracy. The detection accuracy is further improved by introducing the Feature Pyramid Network (FPN) (Lin *et al.*, 2017a). In this network, features are aggregated incrementally from the last layer towards the earlier 5 layers of the CNN network. Afterward, the detection is performed from each of the resulting feature layers using the appropriate anchor boxes.

Parallel to the two-stage detectors, one-stage detectors emerged as faster detectors. However, for a few years, they underperformed in comparison to the two-stage detectors. In the You Look Only Once (YOLO) (Redmon, Divvala, Girshick & Farhadi, 2016), the input image is divided into grid cells. A number of anchor boxes are associated with each grid cell. A CNN is used to extract features and output predictions of each grid cell in one forward propagation. Thus, it is a one-stage detector. This model is good at differentiating objects from the background. However, at the same time, its capacity for precise detection of small objects is limited. The second version of YOLO (Redmon & Farhadi, 2017), included many improvements, notably specifying the anchor boxes based on the distribution of the ground truth bounding boxes in the training data, and pretraining the CNN network for the classification task of high-resolution images in the ImageNet dataset. The third version of YOLO (Redmon & Farhadi, 2018), adopted a feature pyramid-like network where the prediction is performed at three layers and the features are aggregated from the earlier to the last layers. Moreover, they increased the number of layers in the CNN network. Several new versions of the YOLO family have been developed over the years. Encompassing more advanced architectural modifications and training procedures. One-stage detectors suffer greatly from the class imbalance between the objects and the backgrounds. Usually Online Hard Example Mining (OHEM) (Shrivastava, Gupta & Girshick, 2016) is used

to alleviate the class imbalance problem, however; this solution is not enough. Hence Lin, Goyal, Girshick, He & Dollár (2017b) proposed a new classification loss function called the focal loss, this loss is designed to decrease the loss value of the easily classified examples. Consequently, the model focuses more on the difficult ones. The resultant model is referred to as RetinaNet, comprising a ResNet-based CNN integrated with Feature Pyramid Network (FPN), along with dedicated sub-networks for classification and regression tasks. Notably, this model has exhibited superior accuracy compared to prior one- and two-stage detectors, but it was not the fastest. Recently, vision transformer-based models have emerged as a favorable alternative to conventional CNN models in terms of achieving higher accuracy across various tasks. Nevertheless, these models come with a significant computational cost. Furthermore, they lack the inherent bias that CNN models possess, and thus, they require a large training data set. DEtection TRansformer (Detr) (Carion *et al.*, 2020), is one of the early transformer-based detectors. This model contains a combination of a Resnet-based CNN and an encoder-decoder transformer. In Detr, the detection is performed directly using the bipartite matching loss. Therefore, the anchor boxes are not used subsequently the Non-Max Suppression (NMS) is not used either. Consequently, the computational cost of the model is reduced. The detection accuracy of this model is comparable with Faster RCNN.

3.2.1.2 FACE DETECTION

Face detection models share common components with object detection models. Therefore, this section will introduce several renowned face detection models, particularly those that integrate both face detection and facial landmark localization within a multitasking framework. In the Multi-task Cascaded Convolutional Networks (MTCNN) (Xiang & Zhu, 2017), three subnetworks P-Net, R-Net, and O-Net perform face detection and alignment in a cascade manner where the detection and alignment are refined gradually from P-Net to R-Net and finally O-Net. A Non-Max-Suppression is used to select the bounding boxes of higher IoU with the ground truth. This model is trained using an image pyramid where the input image is resized into different scales. This improves the model's ability to detect objects of different sizes. While the

MTCNN performs the detection in three steps, RetinaFace (Deng *et al.*, 2020) detects the face in one step. RetinaFace incorporates many state-of-the-art object detection components. It has a Resnet-based backbone and an FPN. A deformable convolution context module is applied on each level of the FPN followed by a multitask detection head. The latter contains four tasks : face detection, face box regression, facial landmark regression, and dense face regression.

3.2.1.3 BENCHMARK DATASETS

The progression of state-of-the-art detection models has been enabled by the availability of large datasets that serve as references for evaluating the performance of various models. The common Objects in Context COCO2017 (Lin *et al.*, 2014) dataset, is currently the most used benchmark dataset for object detection. The dataset has annotations for general object detection, panoptic segmentation, and other tasks. It contains 118000 training images along with 5000 validation images. The main evaluation criteria for this dataset is the Average Precision at different IoU thresholds. Other datasets have been proposed for specific detection tasks. WiderFace (Yang *et al.*, 2016) is the main face detection benchmark dataset. It consists of over 32000 images with over 393k labeled faces. It has diverse images of faces with different scales, poses, occlusions, expressions, and illuminations. It is partitioned into a 40% training set, a 10% validation set, and a 50% testing set. This dataset has only bounding box annotations. The testing images are categorized into easy, medium, and hard according to the detection accuracy of the EdgeBox (Zitnick & Dollár, 2014). Average precision at an IoU of 0.5 is the evaluation metric used for this benchmark dataset.

3.2.1.4 FACE DETECTION IN THE PRESENCE OF OCCLUSIONS

Occlusion has been one of the main challenges hindering face detection models. Hence, different strategies have been adopted to deal with this problem : First, training on a large dataset that contains many partially occluded faces such as WiderFace ; Second, applying augmentation techniques to simulate occlusion-like images, such as RandomCrop ; Finally, the use of the

attention mechanism in the model architecture to increase the model robustness to occlusion. The facial Attention Network (Wang, Yuan & Yu, 2017), is a RetinaNet face detection-based model with a modified detection head and loss function. In this model, a spatial attention module is added to each level of the feature pyramid network, it has three convolution layers followed by an exponential activation function. The resulting features are element-wise multiplied by input from the current feature pyramid level. The size and the shape of the anchor boxes are specified according to the distribution of the faces in the training data of the WiderFace. These anchor boxes are assigned to each level of the FPN, the smallest anchor boxes for the earlier features and the biggest ones for the features of the last layer. In addition to the classification and regression loss, a pixel-wise attention loss is added to train the attention module. Its ground truth consists of faces that correspond to the relevant feature map level. During training, random crop augmentation is used to crop the input images, this technique allows for the creation of partial faces. Consequently, it increases the model's robustness to occlusion. This model has been tested on WiderFace datasets and it showed a considerable accuracy improvement, especially in the hard set of the data.

3.3 MATERIALS AND METHODS

3.3.1 Data Collection

Upon the approval of the research ethical committee of CHU Sainte-Justine with protocol number 2020-2287 which was approved on February 6, 2020, videos of two minutes along with depth and thermal images of child patients are periodically collected from the PICU of CHUSJ. To record the videos, a camera is placed at the end of the patient's bed to capture the patients and their surroundings. The camera is located approximately one meter from the patient's face. In this study, one RGB image is randomly selected from each of 282 patient videos. Each video is obtained after consent from the patient's guardian. For certain patients, a thermal image of the individual is captured subsequent to the RGB video registration. Subsequently, we obtained 133 thermal images of patients in the PICU.

The inclusion of patients spanning a broad age range from infancy to 18 years old, and incorporating the patient's surroundings, amplifies the effectiveness of our dataset for training computer vision models for the real-world PICU.

3.3.2 Data Labeling

1) Labeling the bounding box for the RGB images :
In the RGB images, the VGG Image Annotator (VIA) software (Dutta & Zisserman, 2019; Dutta, Gupta & Zissermann, 2016) is used to label the faces' bounding boxes of the patients along with the people present in the patients' rooms. Detecting all the people present in the room instead of only the patient can help increase the model's robustness and make it useful for applications that require the presence of the patient along with other people.

2) Labeling the key points for the RGB and thermal images :
A selection of four key points is required from RGB images and their thermal equivalents to estimate the transformation matrix between them. Therefore, the VIA software (Dutta & Zisserman, 2019; Dutta *et al.*, 2016) is used to label four points of corners from objects present in pairs of RGB and thermal images.

3.3.3 Using thermal images to locate the patient region in RGB images

The thermal images were used to roughly estimate the patient location region in the RGB images. This is achieved using the homography matrix when possible and thresholding otherwise.

3.3.3.1 Using the homography matrix

To estimate the parameters of the homography matrix H, a minimum of four corresponding points from both the RGB and thermal images are required. Commonly, more points are estimated using a feature extraction algorithm such as Scale-Invariant Feature Transform (SIFT) (Lowe,

2004); then the locations of similar features are used to estimate the homography matrix. Some of the estimated locations may not be accurate. Subsequently, they are removed using an appropriate algorithm such as Random Sample Consensus (RANSAC) (Fischler & Bolles, 1981). An application of this method can be found in the work of (Negishi *et al.*, 2020). In our work, we manually annotated four points from each image then we estimated the homography matrix.

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix}$$

$$x_{thermo} = \frac{h_{11}x_{RGB} + h_{12}y_{RGB} + h_{13}}{h_{31}x_{RGB} + h_{32}y_{RGB} + h_{33}}, \tag{3.1}$$

$$y_{thermo} = \frac{h_{21}x_{RGB} + h_{22}y_{RGB} + h_{23}}{h_{31}x_{RGB} + h_{32}y_{RGB} + h_{33}}, \tag{3.2}$$

Where x_{thermo}, y_{thermo}, x_{RGB}, y_{RGB} represents the coordinates of the keypoints pair from the thermal and the RGB images respectively. Using the equations 3.1 and 3.2, the values h_{ij} of the homography matrix H can be estimated, then the thermal image can be warped to the corresponding RGB image reference, by multiplying its values with the homography matrix.

3.3.3.2 Using thresholding

For the thermal images where estimating the homography matrix was not possible, because of the absence of sharp corners in RGB-thermal pair images, we performed four steps to locate the patients in the RGB images based on their thermal counterparts, as illustrated in Figure 3.1 : Enum environment test :

1. We divided each thermal image into 16 rectangles.
2. We replaced the pixel values of each rectangle with its maximum pixel value.
3. The rectangles with the highest pixel intensity, indicative of the patient and nearby equipment, are subsequently selected.

4. These selected regions are then matched to corresponding areas within the RGB images, resulting in more patient-focused RGB images.

The thermal and RGB images are not well aligned in space, resulting in a discrepancy between the patient's location in the thermal image and their location in the RGB image. Consequently, many of the resulting RGB images are still large. However, for very small patients, where face detection is hard, the obtained RGB images are small because the corresponding thermal region of the patient is very small.

In the resulting RGB images, if their width is less than 540 pixels or their height is less than 640 pixels, they are padded with zero values. The CropAndPad function of Albumentations (Buslaev *et al.*, 2020) library is used to pad the resulting images with random values of width and height between 100 and 300 pixels. The resulting average heights and width of the RGB images in the processed dataset are respectively 944.79 pixels and 1354.67 pixels compared to 1080 pixels and 1915.35 pixels in the original dataset.

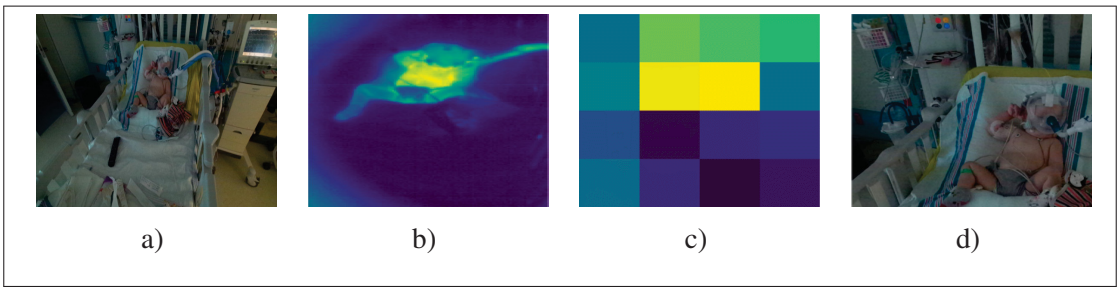

Figure 3.1 Processing the images with a threshold estimated from the pixel intensity of the thermal images. (a) is the initial RGB image of the patient. (b) is the thermal image of the patient child. (c) is the pixel intensity distribution of the thermal image, it is estimated by dividing the thermal image into $4X4$ and filling the rectangular region with the maximum value of that region in the original thermal image. (d) The processed RGB image, after removing the lower and the left regions corresponding to the low intensity in the thermal image

3.3.4 Models

In this section, we present the architecture and the loss functions of the two models we employed in our study.

3.3.4.1 RetinaFace

RetinaFace is one of the widely used and reliable models of face and landmark detection applications. RetinaFace architecture contains four components : Backbone, Feature Pyramid (FP), Context module, and Loss Head. In the original paper, a pre-trained ResNet-152 is used as the Backbone to achieve the highest performance on the hard set of the WiderFace dataset. With 152 layers and an input of 640X640, the resulting features have a large receptive field over the input image which allows it to make better detections. An additional layer is added to the backbone in order to create the FP part of the model by getting the feature map from the last five layers of the backbone. In the feature maps, the features are incrementally aggregated from the last layer with the earlier layers. The resulting feature maps are then fed into a context module to enlarge the receptive field of the model. Finally, a loss head module is applied to the resulting features of the context module to estimate the loss value of the different model tasks. The main advantage of RetinaFace is the use of multitask training with four loss functions : classification loss, bounding box loss, landmarks loss, and dense face regression. In this work, we trained the model for face detection but not for facial landmark localization. Hence, for training, we only included classification and bounding box losses as illustrated in eq 3.3.

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*), \tag{3.3}$$

where L_{cls} is a softmax loss for the face and non-face classes, p_i^* is a binary value of the positive and negative anchors, p_i corresponds to the probability of an anchor i to contain a face $L_{box}(t_i, t_i^*)$ is the box regression loss, $t_i = \{t_x, t_y, t_w, t_h\}_i$ are the coordinates of the predicted bounding box and $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}_i$ are the coordinates of the ground truth bounding box. Detr is a general object detection model created by combining CNN and transformer-based layers,

thus it is a hybrid model. We chose this model instead of a fully transformer-based model to illustrate the advantages of using transformer layers while reducing the computational costs of a comparable fully transformer-based model. In Detr, a Resnet50 backbone is combined with an encoder-decoder transformer model with 6 layers for the encoder and 6 layers for the decoder. The inputs of the encoder are the output features of the backbone added to a positional encoding. Similarly, the decoder's inputs are the feature outputs of the encoder with an added positional encoding. The latter corresponds to the number of the maximum objects to be predicted as they serve as the object queries. A fully connected layer is applied to the output of the decoder to predict the bounding boxes and their corresponding classes. One of the main innovations associated with the Detr model is in the loss function. Instead of using the anchor boxes along with heavy processing to select the appropriate predicted objects of the corresponding ground truth objects, a bipartite matching strategy is used to associate the predicted bounding boxes with their ground truths. Finally, a Hungarian loss function is computed from the selected pairs of the predicted and ground truth values as illustrated in eq 3.4, 3.5.

$$L_H(y, \hat{y}) = \sum_{i=1}^{N} \left[-log\hat{P}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \phi\}} L_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) \right], \tag{3.4}$$

where $\hat{\sigma}(i)$ is the index of the selected prediction, $-log\hat{P}_{\hat{\sigma}(i)}(c_i)$ is the log probability of the class c_i, $L_{box}(b_i, \hat{b}_{\hat{\sigma}(i)})$ is the combination of l_1 loss and the generalized IoU loss.

The bounding box loss is defined as follows :

$$L_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) = \lambda_{iou} L_{iou}(b_i, \hat{b}_{\hat{\sigma}(i)}) + \lambda_{L1} ||b_i + \hat{b}_{\hat{\sigma}(i)}||, \tag{3.5}$$

where $\lambda_{iou}, \lambda_{L1} \in \mathbb{R}$ are hyperparameters.

3.3.5 Evaluation Metrics

Object detection evaluation metrics are extensively analyzed by Padilla, Netto & Da Silva (2020). These metrics are derived from the classification metrics. Accordingly, Precision and Recall are adopted for the detection task.

To estimate precision and recall, True Positive (TP), False Positive (FP), and False Negative (FN) in the detection task are defined as follows :

TP : detected ground truth bounding box.

FP : detected nonexistent ground truth bounding box or misplaced an existing ground truth bounding box.

FN : undetected ground truth bounding box.

TP, FP, and FN are determined by a threshold of the Intersection over Union (IoU) between the predicted bounding box A and the ground truth bounding box B :

$$IoU = \frac{A \cap B}{A \cup B},$$

The Average Precision (AP) is estimated by approximating the Area Under the Curve (AUC) of the accumulated Precision x Recall curve of each class. Finally, the mean average precision of a given IoU threshold is estimated for the entire dataset :

$$mAP_{th} = \frac{1}{N} \sum_{n=1}^{N} AP_n,$$

where N is the number of classes, and th is the threshold of the intersection over union IoU. In this work there is only one class which is the face so N = 1.

3.3.6 Training

3.3.6.1 Training on the WiderFace

Training a transformer is challenging because transformer-based models are sensitive to parameter initialization and optimization learning rate (Touvron *et al.*, 2021). A small change in the learning rate can result in a substantial change in the obtained results. Moreover, it can get stuck in a sharp local minima (Foret, Kleiner, Mobahi & Neyshabur, 2020). Hence, adequate optimization is required to obtain good results. We trained Detr on the WiderFace dataset for 104 epochs with a batch size of 6. Following the original paper of Detr (Carion *et al.*, 2020), we used the AdamW optimizer (Loshchilov & Hutter, 2017) with two learning rates for the CNN and transformer layers. Through careful tuning of the model hyperparameters, we determined the optimal settings, leading us to opt for a learning rate of $5 * 10^{-}5$ for the transformer-based layers, and $5 * 10^{-}6$ for the CNN-based layers, along with a weight decay of $5 * 10^{-}5$. To counter the potential of the model becoming entrenched in sharp local minima, a cosine restart learning rate scheduler (Loshchilov & Hutter, 2016) was adopted, triggering restarts every 12880 iterations. To the best of our knowledge, this is the first study to present the results of training the Detr model on the detection of all faces in images of the Widerface dataset.

3.3.6.2 Fine Tunining on the PICU data

A five-fold cross-validation was used to train the models on both datasets. Hence in each fold, 225 images are used for training and 57 images for validation. Each image presents a different patient. Therefore, the patients in the training and the testing sets are different.
In both the original and the processed PICU datasets, RetinaFace was trained for 50 epochs. During training and similar to (Deng *et al.*, 2020), the images were resized and normalized. Additionally, random brightness contrast and hue saturation value augmentations were applied to the input images.
We conducted training for the Detr model over a span of 50 epochs. Adhering to the methodology outlined by Carion *et al.* (2020), the images underwent random resizing. In this process, the

maximum resize size was set at 800 for both the longest and shortest edges. The training details of both models are shown in Table 3.1.

Tableau 3.1 Training hyperparameters for RetinaFace and Detr on the PICU data

hyperparameters	RetinaFace	Detr
Epochs	50	50
Initial Learning Rate	$lr = 1e-4$, weight decay=$1e-4$	$lr = 5e^{-}5$, weight decay=$5e^{-}5$
Optimizer	SGD	AdamW
Scheduler	Cosine Annealing with Warm Restarts	Step Learning Rate
Processing/Augmentation	Normalize, RandomBrightnessContrast, HueSaturationValue	Normalize, rescale
Train/test	225/57	225/57

3.4 Results and Discussion

3.4.1 Results on WiderFace dataset

At an IoU range of 0.5 to 0.95, the Detr model has a good AP of 61.9% for large faces. In contrast, the validation AP for all faces within the same IoU range is notably lower at 12.4%. The first reason is that Detr is not good at detecting small faces. This is shown in Table 3.2 and illustrated in Figure 3.2. A second possible reason is that Detr detection accuracy is influenced by the number of objects in the input image. As such when there are more than 40 objects in the same image the detection accuracy starts to decline (Carion *et al.*, 2020). Further investigation of the mentioned reasons for subpar validation accuracy on the WiderFace is beyond the scope of this paper. However, the new weights of Detr obtained from training on the WiderFace are useful when the model is applied to the data of the PICU patients.

Figure 3.2 An example of the face detection performance on a WiderFace image at a confidence threshold of 0.9. From this figure, we can see that Detr is not good at detecting small faces

Tableau 3.2 Face Detection accuracy of Detr on the WiderFace

Average Precision	Area	Detr
AP@[IoU=0.5]	All	0.272
AP@[IoU=0.75]	All	0.099
AP@[IoU=0.5 :0.95]	All	0.124
AP@[IoU=0.5 :0.95]	Small	0.025
AP@[IoU=0.5 :0.95]	Medium	0.4
AP@[IoU=0.5 :0.95]	Large	**0.619**

3.4.2 Results of the face detection in PICU

3.4.2.1 Face detection in PICU of both models before fine tuning

Face detection models are commonly evaluated at an IoU threshold of 0.5. Hence in our discussion, we use accuracy to describe the Average Precision (AP) of a model at an IoU of

0.5. As can be seen in Table 3.3, Detr detection accuracy on the original PICU dataset is 11% better than RetinaFace. More importantly, in the processed PICU dataset, where many of the input images are smaller and more focalized on the patient, Detr substantially outperformed RetinaFace by 15.3% as can be seen in Table 3.4. The high performance of Detr on the PICU data

Tableau 3.3 Patient face detection accuracy of both models on Original PICU Dataset without fine tuning

Average Precision	RetinaFace	Detr
AP@[IoU=0.5]	0.531	**0.641**
AP@[IoU=0.75]	**0.311**	0.144
AP@[IoU=0.5 :0.95]	**0.310**	0.249

Tableau 3.4 Patient Face Detection accuracy of both models on the processed PICU Dataset without fine tuning

Average Precision	RetinaFace	Detr
AP@[IoU=0.5]	0.563	**0.716**
AP@[IoU=0.75]	**0.328**	0.197
AP@[IoU=0.5 :0.95]	**0.324**	0.311

after pre-training on the WiderFace shows that this model is able to generalize well compared to RetinaFace. The study of transformer properties such as generalization and robustness to occlusions is a recent active area of research (Caron *et al.*, 2021; Naseer *et al.*, 2021; Zhang *et al.*, 2022). Therefore, the results presented in this work are an important contribution in this context. For both models, the detection accuracy increased after processing which resulted in a reduced detection region. This shows that creating a more patient focused image using a closer camera or zooming in the patient region would allow for a better detection accuracy. Similarly, using thermal images can be helpful to automatically select the patient region from a better detection accuracy.

3.4.2.2 Face Detection in PICU of both models after fine tuning

The detection accuracy of both models RetinaFace and Detr have improved considerably after being fine-tuned on the original dataset, as well as on the processed dataset of the PICU as shown in Table 3.5 and in Table 3.6. At an IoU of 0.5, the RetinaFace model accuracy improved by 7.56% in the original dataset and by 21.1% in the processed dataset. For the same IoU threshold, The detection accuracy of Detr improved by 8.9% in the original dataset and 5.94% in the processed dataset.

Tableau 3.5 Face detection accuracy after fine-tuning of RetinaFace and Detr in both the original and the processed datasets for an AP@[IoU=0.5]

	Original Data		Processed Data	
	RetinaFace	Detr	RetinaFace	Detr
Fold1	0.593	0.731	0.758	0.740
Fold2	0.661	0.737	0.780	0.800
Fold3	0.552	0.667	0.715	0.678
Fold4	0.683	0.825	0.889	0.868
Fold5	0.539	0.690	0.728	0.791
Average	0.606	**0.730**	0.774	**0.775**

Tableau 3.6 The face detection accuracy of RetinaFace and Detr in both the original and the processed datasets for an AP@[IoU=0.5 :0.95]

	Original Data		Processed Data	
	RetinaFace	Detr	RetinaFace	Detr
Fold1	0.303	0.325	0.422	0.388
Fold2	0.312	0.351	0.44	0.392
Fold3	0.281	0.292	0.433	0.310
Fold4	0.370	0.423	0.472	0.437
Fold5	0.283	0.312	0.396	0.342
Average	0.310	**0.341**	**0.433**	0.374

There are two main reasons for the substantial improvement after processing the data as well as after fine-tuning. First, the improvement obtained from fine-tuning can be attributed to the

distribution difference between the clinical data compared to the WiderFace dataset. In the PICU, the patient population is younger, and most of the patients have some occlusion on their faces, some of them have their faces wholly covered with the breathing mask. Second, the original images have high resolution while some patients are very small, hence they occupy a tiny region of the input images. Consequently, it is difficult for RetinaFace and Detr to detect them. Therefore, selecting parts of the images where the patient is located enabled both models to better detect the patients' faces. As mentioned previously, in the case of the patients with small body sizes, the obtained region from the thermal data is smaller than the region obtained from patients with bigger body sizes. Therefore, the equivalent region in the RGB data is more patient-focused. In contrast to the original dataset, RetinaFace AP@50 in the processed dataset is comparable to that of the Detr model. As such, it is clear that RetinaFace with Resnet50 backbone has limited performance in large images. The receptive field of this backbone is limited, hence it is possible that with a deeper backbone such as the Resnet101, the performance on the original dataset can improve. The receptive field problem is less present in transformer-based networks. Thanks to the attention mechanism, these networks are able to model long-range dependencies. Despite being a hybrid model rather than a complete transformer model, Detr has demonstrated impressive performance on the original dataset. Detr exhibits superior accuracy compared to RetinaFace in both scenarios, both before and after fine-tuning on the PICU data. However, as illustrated in Table 3.6, when assessing the models' performance at an IoU range of 0.5 to 0.95, it becomes evident that the fine-tuned RetinaFace significantly outperforms Detr on the processed PICU data. This result highlights the advantage of RetinaFace, particularly in applications that demand a higher level of precision in face detection.

3.4.2.3 Qualitative Analysis

The images from Figure 3.3 confirm the quantitative results. The RetinaFace model has a low performance on the original images but has a good performance on the processed ones. Detr showed a better performance on both datasets, but interestingly the processed dataset has led to more false positives for Detr. We can also see that the bounding box of Detr is less accurate than

Figure 3.3 The detection accuracy before fine-tuning. Some of the images in the original data are small because they are cut to avoid identifying the patients. Some of the images in the processed data were not modified because there were no thermal equivalents to them or their thermal equivalent were substantially misaligned with the RBG images, hence it was not possible to use the thermal images to obtain the region of the patient

Figure 3.4 The detection accuracy in the fold5 after fine-tuning. Some of the images in the original data are small because they are cut to avoid identifying the patients. Some of the images in the processed data were not modified because there were no thermal equivalent to them or their thermal equivalent were substantially misaligned with the RBG images, hence it was not possible to use the thermal images to obtain the region of the patient

RetinaFace. Moreover, both models have shown a good performance in the presence of occlusion. Remarkably, Detr has shown superior detection capabilities even when the entire face is obscured.

Figure 3.4 shows the results of both models after fine-tuning. RetinaFace model was able to detect relatively easy faces in the original images. However, it struggled to detect more challenging faces with severe occlusion or extreme pose. On the other hand, the Detr model did not substantially improve when it comes to the false positives problem and detection accuracy. Detr showed a superior performance in detecting faces with extreme occlusion. Its ability to model contextual information is very useful in detecting difficult faces. However, it makes the model prone to false positives. This can be illustrated by instances where even the face of a teddy bear located on the patient's bed triggered face detection.

3.5 Conclusion

The current state-of-the-art CNN-based face detection models perform very well in public datasets such as the WiderFace. However, such high performance may not translate well to all real-world applications. For instance, the face detection accuracy of RetinaFace in the PICU is lower than reported in the WiderFace. This is due to the challenges present in the PICU. Mainly, occlusion, extreme pose, and the presence of face-like shapes that mislead the model detection. In this study, we showed that the transformer-based detection model Detr, generalizes very well to the PICU data after being pre-trained on the WiderFace dataset, despite the distribution difference between the two datasets. Therefore, in the absence of large datasets, such as in applications related to the clinical setting ; a pre-trained Detr may perform better than non-transformer-based models. Moreover, our experiments showed that in detecting faces in large images, Detr performed better in accuracy than RetinaFace with a Resnet50 backbone.
Despite the relatively good performance of the two models discussed in this study, it is important to acknowledge certain limitations they exhibit. First, Detr suffers greatly from false positives where face-like shapes are detected as actual faces. This problem can be solved implicitly by increasing the amount of the PICU training data, or by developing a tailored solution that takes

into account the unique characteristics of the PICU environment, including aspects such as facial area, location, and patient movements. RetinaFace on the other hand, with a Resnet50 backbone is limited in its ability to detect faces in big images. Bigger backbone networks may yield better results, but they are computationally more expensive.

In the current study, we did not investigate the speed of the developed models. Transformer-based models are generally slower than CNN-based ones. However, thanks to its direct detection mechanism, the detection speed of Detr is comparable to the CNN-based ones (Carion *et al.*, 2020).

CHAPITRE 4

DISCUSSION

La généralisation des modèles d'apprentissage profond est actuellement un domaine de recherche actif. Cette propriété de généralisation n'est pas seulement utile pour les tâches avec des données limitées, mais aussi pour les applications du monde réel présentant un décalage dans la distribution des données. Jusqu'à récemment, les réseaux de neurones convolutionnels étaient généralement considérés comme les meilleurs dans les applications de vision, y compris leur capacité à généraliser par rapport aux modèles classiques d'apprentissage automatique. La généralisation des modèles CNN est davantage liée à la diversité des données sur lesquelles ils sont entraînés et aux approches d'entraînement utilisées pour ces modèles, plutôt qu'aux propriétés inhérentes des modèles CNN. L'invention de l'architecture de Transformer a permis le développement de modèles plus performants avec une capacité de généralisation encore meilleure. La capacité de généralisation de ces nouveaux modèles de Transformer développés est davantage liée à leur architecture qu'à d'autres facteurs.

Certains avancent que les modèles de vision transformer apprennent les formes à partir de l'image d'entrée plutôt que les textures apprises par leurs homologues CNN. Cependant, une compréhension complète des raisons derrière les propriétés intrigantes des modèles Transformer de vision reste à accomplir, c'est pourquoi des recherches sont en cours pour examiner comment ces propriétés se produisent pendant l'entraînement d'un modèle de Transformer de vision. Dans notre étude, nous avons proposé d'exploiter la propriété de généralisation des Transformers pour créer un meilleur modèle de détection de visage pour l'environnement clinique. Plus précisément, nous avons entraîné le modèle de détection d'objets Detr sur l'ensemble de données public WiderFace, composé de plus de 13 000 images contenant plusieurs visages dans différents contextes, puis nous avons testé le modèle entraîné sur des images des patients en soins intensifs pédiatriques (USIP).

Les résultats obtenus ont montré que le modèle basé sur le Transformer, Detr, surpassait largement RetinaFace. Cela valide notre hypothèse selon laquelle le modèle basé sur l'architecture de Transformer peut généraliser mieux que les autres modèles. Ce résultat est bénéfique non seulement pour la détection de visages en environnement clinique, mais il ouvre la voie à

davantage d'applications en vision où les données sont limitées. De plus, une fois que le modèle basé sur le Transformer est déployé dans l'environnement de production, il peut être plus robuste face à un changement dans la distribution des données d'entrée par rapport aux modèles précédents. Ce changement de distribution peut survenir pour de nombreuses raisons, telles qu'un changement d'éclairage, une modification de la structure de la salle du patient, un changement de vêtements du patient, etc.

L'architecture du modèle Transformer lui permet d'extraire les caractéristiques de l'ensemble de l'image d'entrée dès sa première couche. En revanche, dans les modèles CNN, les caractéristiques sont progressivement extraites de la première couche à la dernière couche, de sorte que leur champ réceptif est limité par le nombre de couches du modèle (profondeur). Par conséquent, Les modèles basés sur Transformer apprennent mieux le contexte de l'ensemble de l'image, ce qui les rend plus robustes à l'occlusion et performants sur les grandes images par rapport aux modèles CNN. Par exemple, après avoir affiné le modèle Detr sur l'ensemble de données original des patients en unité de soins intensifs pédiatriques (USIP), il a surpassé le modèle RetinaFace affiné. Cependant, après le traitement de l'ensemble de données, les dimensions des images sont devenues plus petites, ce qui a entraîné des performances comparables en termes de précision entre les deux modèles Detr et RetinaFace. À un chevauchement plus élevé entre la boîte englobante de la vérité terrain et la boîte englobante prédite (Intersection over Union, IoU), RetinaFace a généralement de meilleures performances de détection que Detr, que ce soit sur l'ensemble de données original ou sur l'ensemble de données traité. Cela peut être attribué au processus de détection de RetinaFace, qui repose sur des boîtes d'ancrage préalablement proposées au lieu de détecter directement les visages présents dans l'image. L'utilisation des images thermiques a contribué à produire des images plus petites, davantage axées sur les patients, ce qui a entraîné de meilleures performances de détection de visage pour les deux modèles. Cela montre l'importance de tirer parti de différentes modalités pour améliorer la précision de la détection.

CONCLUSION ET RECOMMANDATIONS

Le modèle de détection de visage basé sur le Transformer présenté dans cette étude a mis en évidence le potentiel de généralisation de ce modèle et ses avantages par rapport au modèle basé sur CNN, RetinaFace. Les modèles basés sur le Transformer ont une meilleure capacité de généralisation, ce qui en fait une option prometteuse pour développer des solutions basées sur l'IA dans des scénarios où les données sont limitées, ce qui est fréquent dans le contexte clinique. Comme nous l'avons discuté précédemment, les modèles de vision basés sur le Transformer possèdent des propriétés remarquables. Ces propriétés peuvent être exploitées dans des applications de vision au-delà de la détection de visage. Les avantages de ces modèles résident dans la qualité des représentations qu'ils apprennent lors de l'entraînement. Par conséquent, il est donc possible de concevoir de manière créative des tâches auxiliaires qui permettent d'apprendre des représentations extrêmement utiles pour des tâches ultérieures, même lorsque très peu d'exemples d'entraînement sont disponibles, ce qui est particulièrement pertinent dans le domaine clinique. Le modèle développé dans cette étude peut être employé pour la détection des visages des patients dans l'unité de soins intensifs pédiatriques. En général, Detr a démontré une meilleure capacité à s'adapter aux changements de distribution. Ces changements peuvent survenir en clinique, en raison de facteurs tels que le changement dans la démographie des patients, les conditions d'éclairage, l'équipement connecté aux patients, l'aménagement de la chambre, et d'autres facteurs. Cependant, dans des applications où une détection de visage de patient avec une très haute précision est nécessaire, RetinaFace offre une meilleure précision. Il est à noter que les deux modèles de notre étude ont montré de bonnes performances en présence d'une occlusion modérée. Cependant, ils ont encore besoin d'améliorations pour détecter les visages avec une occlusion substantielle, et cela pourrait être réalisé en collectant davantage de données, éventuellement en les générant synthétiquement. Il est également important de noter que Detr est un modèle hybride, combinant des couches CNN et Transformer. Il est possible qu'un modèle entièrement basé sur le Transformer, comme YOLOS, puisse surpasser à la fois Detr

et RetinaFace en termes de performance, mais cela pourrait nécessiter davantage de puissance de calcul pour l'entraînement. Le modèle de détection de visage du patient développé dans ce travail peut être intégré dans un pipeline d'une application d'évaluation automatique de l'état du patient, qui repose sur la région faciale. Cette application, telle que l'évaluation de la conscience, de la douleur et de la détresse, nécessite également un modèle entraîné sur de grandes quantités de données. Par conséquent, il pourrait être bénéfique d'exploiter les propriétés du Transformer pour développer des applications fiables dans un environnement clinique contraint. Étant donné que les états du patient peuvent être évalués à partir de son visage, le modèle développé pourrait être conçu comme un modèle multitâche. Dans un modèle multitâche, un réseau d'extraction de caractéristiques peut être combiné avec différents sous-réseaux, chacun étant responsable d'un état spécifique du patient, tel que la conscience, la douleur et la détresse. Le réseau d'extraction de caractéristiques pourrait être pré-entraîné pour la reconnaissance faciale, où le modèle apprendrait les caractéristiques des expressions faciales, des petits mouvements de tête et des points de repère faciaux. Ces caractéristiques pourraient également être apprises à l'aide de techniques d'auto-apprentissage.

BIBLIOGRAPHIE

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T. & Veit, A. (2021). Understanding robustness of transformers for image classification. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M. & Kalinin, A. A. (2020). Albumentations : fast and flexible image augmentations. *Information*, 11(2), 125.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. (2020). End-to-end object detection with transformers. *European conference on computer vision*, pp. 213–229.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660.

Cavanna, A. E., Shah, S., Eddy, C. M., Williams, A. & Rickards, H. (2011). Consciousness : a neurological perspective. *Behavioural neurology*, 24(1), 107–116.

Cologan, V., Schabus, M., Ledoux, D., Moonen, G., Maquet, P. & Laureys, S. (2010). Sleep in disorders of consciousness. *Sleep medicine reviews*, 14(2), 97–105.

Deng, J., Guo, J., Ververas, E., Kotsia, I. & Zafeiriou, S. (2020). Retinaface : Single-shot multi-level face localisation in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*.

Dosso, Y. S., Kyrollos, D., Greenwood, K. J., Harrold, J. & Green, J. R. (2022). NICUface : Robust neonatal face detection in complex NICU scenes. *IEEE Access*, 10, 62893–62909.

Dutta, A., Gupta, A. & Zissermann, A. [Version : X.Y.Z, Accessed : [Accessed : March 4, 2023]]. (2016). VGG Image Annotator (VIA). Repéré à http://www.robots.ox.ac.uk/~vgg/software/via/.

Dutta, A. & Zisserman, A. (2019). The VIA annotation software for images, audio and video. *Proceedings of the 27th ACM international conference on multimedia*, pp. 2276–2279.

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Foret, P., Kleiner, A., Mobahi, H. & Neyshabur, B. (2020). Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv :2010.01412*.

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

Gleichauf, J., Hennemann, L., Fahlbusch, F. B., Hofmann, O., Niebler, C. & Koelpin, A. (2023). Sensor Fusion for the Robust Detection of Facial Regions of Neonates Using Neural Networks. *Sensors*, 23(10), 4910.

Gowtham. (2018, Apr). Deep learning-our shot towards real ai. Medium. Repéré à https://gowtham-palani.medium.com/deep-learning-our-shot-towards-real-ai-c27a5766081f.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1), 87–110.

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 1904–1916.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), 18–28.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Ista, E., van Dijk, M., Tibboel, D. & de Hoog, M. (2005). Assessment of sedation levels in pediatric intensive care patients can be improved by using the COMFORT “behavior” scale. *Pediatric Critical Care Medicine*, 6(1), 58–63.

Johansson, M. & Kokinsky, E. (2009). The COMFORT behavioural scale and the modified FLACC scale in paediatric intensive care. *Nursing in critical care*, 14(3), 122–130.

Kazemi, V. & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874.

Kyrollos, D. G., Greenwood, K., Harrold, J. & Green, J. R. (2022). Transfer learning approaches for neonate head localization from pressure images. *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6.

Li, C., Pourtaherian, A., Van Onzenoort, L., a Ten, W. E. T. & de With, P. H. (2020). Infant monitoring system for real-time and remote discomfort detection. *IEEE Transactions on Consumer Electronics*, 66(4), 336–345.

Lienhart, R., Kuranov, A. & Pisarevsky, V. (2003). Empirical analysis of detection cascades of boosted classifiers for rapid object detection. *Pattern Recognition : 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pp. 297–304.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft coco : Common objects in context. *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017a). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017b). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liuzzi, P., Campagnini, S., Hakiki, B., Burali, R., Scarpino, M., Macchi, C., Cecchi, F., Mannini, A. & Grippo, A. (2023a). Heart rate variability for the evaluation of patients with disorders of consciousness. *Clinical Neurophysiology*, 150, 31–39.

Liuzzi, P., Grippo, A., Draghi, F., Hakiki, B., Macchi, C., Cecchi, F. & Mannini, A. (2023b). Can Respiration Complexity Help the Diagnosis of Disorders of Consciousness in Rehabilitation ? *Diagnostics*, 13(3), 507.

Loshchilov, I. & Hutter, F. (2016). Sgdr : Stochastic gradient descent with warm restarts. *arXiv preprint arXiv :1608.03983*.

Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv :1711.05101*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60, 91–110.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F. & Yang, M.-H. (2021). Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 23296–23308.

Negishi, T., Abe, S., Matsui, T., Liu, H., Kurosawa, M., Kirimoto, T. & Sun, G. (2020). Contactless vital signs measurement system using RGB-thermal image sensors and its clinical screening test on patients with seasonal influenza. *Sensors*, 20(8), 2171.

Padilla, R., Netto, S. L. & Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. *2020 international conference on systems, signals and image processing (IWSSIP)*, pp. 237–242.

Qi, D., Tan, W., Yao, Q. & Liu, J. (2022). YOLO5Face : why reinventing a face detector. *European Conference on Computer Vision*, pp. 228–244.

Ray, P. P. (2023). ChatGPT : A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Redmon, J. & Farhadi, A. (2017). YOLO9000 : better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271.

Redmon, J. & Farhadi, A. (2018). Yolov3 : An incremental improvement. *arXiv preprint arXiv :1804.02767*.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016). You only look once : Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

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.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv :1706.05098*.

Shrivastava, A., Gupta, A. & Girshick, R. (2016). Training region-based object detectors with online hard example mining. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*.

Sternbach, G. L. (2000). The Glasgow coma scale. *The Journal of emergency medicine*, 19(1), 67–71.

Sun, Y., Shan, C., Tan, T., Tong, T., Wang, W., Pourtaherian, A. et al. (2019). Detecting discomfort in infants through facial expressions. *Physiological measurement*, 40(11), 115006.

Sury, M. R. & Bould, M. D. (2011). Defining awakening from anesthesia in infants : a narrative review of published descriptions and scales of behavior. *Pediatric Anesthesia*, 21(4), 364–372.

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.

Ting, W. K.-C., Perez Velazquez, J. L. & Cusimano, M. D. (2014). Eye movement measurement in diagnostic assessment of disorders of consciousness. *Frontiers in neurology*, 5, 137.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International conference on machine learning*, pp. 10347–10357.

Uijlings, J. R., Van De Sande, K. E., Gevers, T. & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104, 154–171.

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.

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, 1, I–I.

Wang, J., Yuan, Y. & Yu, G. (2017). Face attention network : An effective face detector for the occluded faces. *arXiv preprint arXiv :1711.07246*.

Wannez, S., Hoyoux, T., Langohr, T., Bodart, O., Martial, C., Wertz, J., Chatelle, C., Verly, J. G. & Laureys, S. (2017). Objective assessment of visual pursuit in patients with disorders of consciousness : an exploratory study. *Journal of neurology*, 264, 928–937.

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S. & Traue, H. C. (2016). Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3), 286–299.

Xiang, J. & Zhu, G. (2017). Joint face detection and facial expression recognition with MTCNN. *2017 4th international conference on information science and control engineering (ICISCE)*, pp. 424–427.

Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. (2018). Convolutional neural networks : an overview and application in radiology. *Insights into imaging*, 9, 611–629.

Yang, F., He, S., Sadanand, S., Yusuf, A. & Bolic, M. (2022). Contactless measurement of vital signs using thermal and RGB cameras : A study of COVID 19-related health monitoring. *Sensors*, 22(2), 627.

Yang, S., Luo, P., Loy, C.-C. & Tang, X. (2016). Wider face : A face detection benchmark. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533.

Zamzmi, G., Pai, C.-Y., Goldgof, D., Kasturi, R., Ashmeade, T. & Sun, Y. (2016). An approach for automated multimodal analysis of infants' pain. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 4148–4153.

Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Liu, X. & Liu, Z. (2022). Delving deep into the generalization of vision transformers under distribution shifts. *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 7277–7286.

Zhu, Y., Cai, H., Zhang, S., Wang, C. & Xiong, Y. (2020). Tinaface : Strong but simple baseline for face detection. *arXiv preprint arXiv :2011.13183*.

Zitnick, C. L. & Dollár, P. (2014). Edge boxes : Locating object proposals from edges. *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 391–405.

Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. (2023). Object detection in 20 years : A survey. *Proceedings of the IEEE*.