IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 12, 2022, accepted May 29, 2022, date of publication June 8, 2022, date of current version June 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3181167

NICUface: Robust Neonatal Face Detection in Complex NICU Scenes

YASMINA SOULEY DOSSO[1], (Student Member, IEEE), DANIEL KYROLLOS[1], (Student Member, IEEE), KIMBERLEY J. GREENWOOD[2,3], JOANN HARROLD[4], AND JAMES R. GREEN[1], (Senior Member, IEEE)

[1]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada
[2]Department of Mechanical Engineering, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[3]Department of Clinical Engineering, Children's Hospital of Eastern Ontario, Ottawa, ON K1H 8L1, Canada
[4]Department of Neonatology, Children's Hospital of Eastern Ontario, Ottawa, ON K1H 8L1, Canada

Corresponding author: James R. Green (jrgreen@sce.carleton.ca)

This work was supported in part by the IBM Center for Advanced Studies, and in part by the Natural Sciences and Engineering Research Council under Grant CRDPJ 543940-19.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Board of Carleton University under Application No. CU-117311 and CU-107193.

ABSTRACT The development of non-contact patient monitoring applications for the neonatal intensive care unit (NICU) is an active research area, particularly in facial video analysis. Recent studies have used facial video data to estimate vital signs, assess pain from facial expression, differentiate sleep-wake status, detect jaundice, and in face recognition. These applications depend on an accurate definition of the patient's face as a region of interest (ROI). Most studies have required manual ROI definition, while others have leveraged automated face detectors developed for adult patients, without systematic validation for the neonatal population. To overcome these issues, this paper first evaluates the state-of-the-art in face detection in the NICU setting. Finding that such methods often fail in complex NICU environments, we demonstrate how fine-tuning can increase neonatal face detector robustness, resulting in our NICUface models. A large and diverse neonatal dataset was gathered from actual patients admitted to the NICU across three studies and gold standard face annotations were completed. In comparison to state-of-the-art face detectors, our NICUface models address NICU-specific challenges such as ongoing clinical intervention, phototherapy lighting, occlusions from hospital equipment, etc. These analyses culminate in the creation of robust NICUface detectors with improvements on our most challenging neonatal dataset of +36.14, +35.86, and +32.19 in AP30, AP50, and mAP respectively, relative to state-of-the-art CE-CLM, MTCNN, img2pose, RetinaFace, and YOLO5Face models. Face orientation estimation is also addressed, leading to an accuracy of 99.45%. Fine-tuned NICUface models, gold-standard face annotation data, and the face orientation estimation method are also released here.

INDEX TERMS Face detection, neonatal dataset, NICU, complex care scenes, convolutional neural networks, face orientation.

I. INTRODUCTION

Many neonatal non-contact monitoring approaches utilize the patient's facial area as a region of interest (ROI) for diverse tasks including estimation of patient heart rate (HR) or respiration rate (RR) [1]–[6], assessment of pain from facial expression [7]–[9], detection of jaundice [10], [11], sleep-wake detection from facial expression analysis [12], [13], or face recognition to prevent baby swapping or abductions [14], [15]. In these studies, facial ROI selection is often a manual or semi-automated process (*e.g.*, [1]–[3], [5]), or relies on face detection methods developed for adult faces (*e.g.*, [6]) that have not been validated on neonates, especially in complex care scenes.

When estimating HR from facial video data, Fernando *et al.* [1] manually extracted the patient's face to

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

track the changes in skin pixels using adaptive bandpass filtering and principal component analysis. Klaessens *et al.* [2] extracted regions of the face for subsequent HR estimation using the Eulerian video magnification (EVM) technique [16]; although it is not discussed in detail, it appears that facial ROI were manually selected. Kyrollos *et al.* [6] also used EVM on facial video data, but for RR estimation. They leveraged the RetinaNet model [17] for automatic face detection. However, the automatic face detection was limited to within-patient testing and therefore the generalizability of their model is untested. Villarroel *et al.* [3] automatically detected video segments where patient skin is visible and manually identified a ROI (face, head, or neck) for subsequent HR estimation using independent component analysis. They later trained a multi-task convolutional neural network (CNN) for patient detection and skin segmentation to automatically detect the patient and all visible skin area for HR estimation [4]. Since they only performed skin detection, it is unclear how face detection would perform for their application given the varying amounts of visible skin during occlusions from beddings or hospital equipment.

More recently, Huang *et al.* [5] manually detected the patient's face in video recordings for HR estimation. They discussed how challenging and inaccurate an automated face detector would be, considering the variations in patient posture and camera perspectives. Hence, a manual approach was used on the first video frame and a tracking algorithm would perform detection on subsequent frames. Other studies have adopted this approach (manual ROI detection followed by automated tracking) for continuous video-based face detection [18]. This approach can be reliable for short videos [5], or with robust tracking systems [18]; however, laborious manual ROI definition must be repeated at the start of each video and when tracking fails due to occlusions or excessive patient motion.

Khanam *et al.* [19] aimed to overcome this manual task by training a neonatal face detector as a preprocessing ROI detection step for subsequent HR/RR estimation based on colour and motion variations, respectively. They discussed the challenges faced in utilizing a state-of-the-art face detection model due to occlusions, baby poses, and complex hospital settings. Moreover, these models are pretrained on adult populations, often including only a few or no baby images. They then leveraged 473 images collected from online available sources to finetune the YOLOv3 model [20] for neonatal face detection. They unfortunately did not report the performance of their face detector preprocessing step, focusing instead on evaluating the proposed HR/RR methods. Khanam *et al.* discussed the need to obtain a larger neonatal image dataset to improve their method's reliability.

Neonatal facial expression recognition is another important task for assessing patient status. Lin *et al.* [21] aimed to recognize when a 0-2 year old infant is happy, sad, or normal, directly from facial video data. Data were collected by asking parents to capture and submit images and videos of their children using smartphones. For videos, one image was extracted every 30 frames and a similarity matching algorithm (SSIM) [22] was used to remove near-identical images. As an image preprocessing step to standardize the face orientation, they used the Dlib and OpenCV visual libraries for face detection and cropping where the image was rotated at 90°, 180°, and 270° until a face was detected. They however did not report the performance of their face detector nor of their rotation experiment.

Several studies have also examined patient faces for neonatal pain assessment [7]–[9]. Brahnam *et al.* [8] detected painful events from images of swaddled newborns. Images were carefully preprocessed to extract the facial area (*i.e.* images were rotated for standardized face orientation, images were cropped to only include the patient's face). It is unclear if these preprocessing steps were performed programmatically or manually. They later automatically detected faces from video recordings using a Discriminative Response Map Fitting (DRMF) model [23] to analyze the temporal pattern of facial expression during painful procedures [9]. In this latter study, video recordings of occlusions from moving limbs were included. No results from the performance of the face detection step were reported.

Salekin *et al.* [7] also used video recordings for pain assessment evaluated through facial expression, body movement, and crying sound analysis. They leveraged the pretrained YOLOv3 model for face and body detection as a preprocessing step before analyzing facial expression and body motion. No results on the performance of the face/body detectors were reported; they directly evaluated the pain assessment methods from the face, body, and sound data streams. These pain detection studies all share common key points: face detection as a preprocessing ROI detection step followed by feature extraction for facial expression estimation; implementation using a dataset of newborns recorded at a close distance such that the face fills the majority of the frame, with minimal occlusions, and no dark environment.

Neonatal face detection has been applied for jaundice detection that seeks to quantify the yellowing of the skin [10], [11]. These studies use image processing approaches based on the YCbCr color space for skin segmentation [10], followed by a manual ROI detection step to identify a specific facial region [11].

Several recent studies have used face video analysis for sleep-wake state detection in neonates based on their facial expression or opening of the eyes [12], [13], [24]. To detect faces, Mukai *et al.* [12] first rotated images to standardize the face orientation pointing North (unclear if done programmatically or manually), then used the OpenFace library to detect and align faces [25]. They did not report performance of the face detection step; however, they discussed how facial occlusion from bed sheets and lighting variations impeded the accuracy of the detected face, and thereby impeded the classification of sleep-wake cycles. They discussed how neonatal face detection is a difficult task that requires additional research and development.

FIGURE 1. Face detection difficulty from Complex NICU Scenes.

In comparison to relatively standard cameras used in previously mentioned studies, Awais *et al.* [13] recorded patients with the Fluke TiX580 camera which can capture multiple color palettes. To that end, they leveraged the camera's unique specifications by detecting faces based on pixel intensities in the CIELAB (Commission Internationale de l'eclairage, L*a*b*) color space. Studies relying on pixel-intensity based approaches for face detection can be useful for specific neonatal monitoring applications; however, skin-tone-based face detection performance can be limited for dark skin patients, different lighting conditions, or for images including the patient's full body and bed environment (as opposed to a close up facial view).

Another sleep monitoring study was recently conducted on baby manikins by Khan [24], where they created a smart home baby monitor device that detects sleeping postures and notifies the caregiver. Four different events were detected including facial coverage due to prone position, patient removing the blanket, frequent motion, and awake detection. For the latter, they used the Multi-Task Cascaded Neural Network (MTCNN) to detect the face, regression trees to detect a 68-point facial landmarks including 6 landmarks surrounding the eyes, and computed the eye aspect ratio to detect when the eyes are continuously open (*i.e.* baby is awake). Facial coverage due to prone position was detected from nose detection; however, they did not use the MTCNN and regression tree approach since this technique is severely impacted by facial occlusions. Instead, they opted for a pose detection model made of a body skeleton connected to facial landmarks. A prone position is determined by the absence of the connected facial landmarks. Their study was entirely trained and largely tested on baby manikins. They did obtain a few additional images of real babies and infants collected from online available sources to test their methods. Although results were promising on these few images, they discussed how new challenges may arise from real babies with varying facial occlusions, or more complex sleeping poses. The authors emphasized the need to collect such a challenging real-life dataset, to acquire reliable labelled data, and to retrain a pretrained model accordingly.

Neonatal face recognition applications are evaluating newborns' facial features to properly identify patients in hospital as a prevention measure to baby swapping or abduction. Bharadwaj *et al.* [15] performed manual face detection given that existing detectors failed to identify newborn faces. To overcome this issue, Awais *et al.* [14] leveraged the color palettes from the Fluke TiX580 camera for automatic face detection and reported an accuracy of 98.5%. Their dataset used controlled head movements ($-45°$ to $45°$ in yaw head tilt), close camera distance (0.25-0.36 m), and excluded occlusions from limb movements to obtain best quality data for face recognition. The present paper explores a different task by systematically evaluating face detection in a variety of complex NICU scenes for diverse neonatal monitoring studies.

Among all the video-based neonatal monitoring research, numerous approaches to face detection were adopted as a preprocessing step for their specific application using different NICU datasets. It is unclear if the various proposed methods would be reliable in a complex clinical setting when the camera is sometimes placed far from the patient, where the face may be occluded due to ventilation support or other reason, lighting conditions vary widely and change frequently, patient pose can vary, and the scene may capture ongoing clinical interventions. Some of these complex scenes are depicted in Fig. 1. In many previous studies, these challenges are acknowledged and a manual ROI detection method was adopted (*e.g.*, [1]–[3], [5], [8], [11], [18]). In other cases, studies used a state-of-the-art model pretrained on an adult population for ROI detection without rigorous validation on neonatal patients in complex NICU environments (*e.g.*, [6], [7], [9], [12], [13]). In rarer cases, studies have trained a facial ROI detector using neonatal datasets including only a few of the mentioned challenges to obtain a more reliable ROI for their specific neonatal application [14], [19]. It is however unclear if such models are robust and generalizable since extensive evaluations of

such detectors under varying NICU conditions have not been reported.

These important limitations in neonatal face detection motivates the current study in which we (1) rigorously determine conditions where pretrained state of the art face detection models perform accurately and where they fail in complex NICU scenes, and (2) create an improved neonatal face detection model robust to these identified challenges using transfer learning. This paper addresses these needs through the following contributions:

1) Demonstrated limitations of the state-of-the-art pretrained face detection models for neonatal face detection during realistic patient monitoring conditions by utilizing three different neonatal studies presenting different patients and environments.
2) Created two neonatal face detection models (NICUface) by finetuning the most performant pretrained face detection models on exceptionally challenging NICU scenes.
3) Proposed a simple but reliable face orientation estimation approach as a required preprocessing step in neonatal face analysis applications.
4) Provided high quality face annotations for two publicly available benchmark neonatal datasets to promote continued development of the state of the art in neonatal face detection.

This study not only explores the limits of state-of-the-art face detection models, but also overcomes the limitations of neonatal face detection in a clinical setting. A plethora of non-contact neonatal monitoring applications will likely benefit from the robust NICUface detectors presented here.

II. BACKGROUND

A. OBJECT DETECTION

In the last few years, object detection models have improved in both speed and accuracy. Notably, as part of state-of-the-art two-stage object detectors, the R-CNN family has seen various versions including the R-CNN [26], Fast R-CNN [27], and Faster R-CNN [28]. Each edition demonstrates advances in implementing a CNN where region proposal methods suggest areas of the image where an object of interest is suspected to reside, followed by object localization using bounding box regression.

Instead of relying only on selected proposed regions of the image, the You Only Look Once (YOLO) family of object detectors looks at the entire image and simultaneously generates class probabilities within each predicted bounding box [20], [29], [30]. The object of interest corresponds to the highest probability region, thus only requiring to ''look once'' at the image before making a prediction. Such one-stage object detectors have gained popularity due to their fast computation, especially in real-time applications. Redmon *et al.* created three versions of this YOLO architecture from 2015 to 2018, by incrementally improving the model's speed and accuracy [20], [29], [30]. In the past couple of years, other researchers have extended Redmon's work to achieve even better and faster real-time performance with YOLOv4 [31] by using ''bag-of-freebies'' (methods used during training) and ''bag-of-specials'' (post-processing methods used during inference). Among them, significant detection improvement were noticed using a new mosaic data augmentation which creates a tile of four training images thereby helping the model detect small objects while reducing the required mini-batch size during training. Compared the mean square error (MSE) used in YOLOv3 for bounding box regression, YOLOv4 uses a complete IOU (CIOU) loss which compares the predicted and ground truth bounding boxes area by considering the distance between each center points and aspect ratio, in addition to evaluating their overlap from traditional IOU. Compared to the four previous versions, Glenn Jocher [32] introduced YOLOv5 implemented on PyTorch instead of Darknet framework, thereby allowing the implementation of models of various sizes including small and lightweight ones for easy deployment to mobile devices. YOLOv5 also introduces a Focus Layer made up of YOLOv3's first three layers to reduce layers, parameters, and CUDA memory, while improving speed during forward propagation and backpropagation. Overall, YOLOv5 is fastest, more lightweight, and more accurate among the entire YOLO family.

Other prominent recent object detectors include the single-stage object detector RetinaNet which introduces a new Focal Loss optimization that focuses on extreme foreground-background class imbalance during training [17], the EfficientDet [33] model that uses the EfficientNet [34] classifier as a backbone for model scaling, and the DEtection TRansformer (DETR) network that leverages a CNN and transformer encoder-decoder architecture to perform end-to-end object detection with bipartite matching for generating direct predictions [35].

To train these above-mentioned detectors, research groups have often relied on the PASCAL VOC dataset [36] and/or the COCO dataset [37]. These two object detection benchmark datasets were created for various object recognition challenges including classes such as person, cat, bicycle, etc.

B. FACE DETECTION

Detecting the facial area is often performed in three different ways: the detection of the entire face enclosed within a bounding box (face detection), the detection of the geometric structure of the face outlined by specific landmarks (face alignment), or the detection every pixel pertaining to the person's face (face segmentation). All of these applications are depicted in Fig. 2. Facial alignment is typically applied using 5-point landmarks including the center of left eye, center of right eye, tip of nose, left corner of mouth, and right corner of mouth [38], [39]. In other cases, finer facial structure is extracted with 68-point landmarks including eyebrow line, eye contour, length and width of nose, upper and lower lip contour, and jawline [40]. In face segmentation, the whole face is either segmented as a whole [41] or is segregated into different facial regions (*e.g.*, eyes, nose,

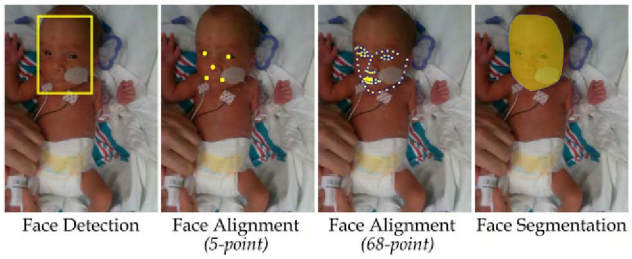

FIGURE 2. Face detection techniques.

mouth, skin, hair) [42]. Face alignment and segmentation are particularly useful in further facial analysis applications such as face recognition or facial expression detection; however, they are more difficult tasks to achieve compared to detecting bounding boxes. Only face detection results from bounding box predictions are investigated quantitatively in this study, while facial alignment methods are evaluated qualitatively.

1) BENCHMARK DATASETS

To train and evaluate face detection models, several benchmark face image datasets are available.

Face detection benchmark datasets include:

- WIDER FACE [43]: Images include faces with variations in scale, pose, occlusion, expression, makeup, and illumination (393,703 annotated faces from 32,203 images). Different subsets are included as Easy, Medium, and Hard data, based on the increasing level of difficulty to detect the face due to varying scale, occlusion, and pose.
- FDDB [44] (Face Detection Dataset and Benchmark): Images including faces with variations in occlusions, poses, resolution, and out-of-focus faces (5,171 annotated faces from 2,845 images).

Face alignment benchmark datasets include:

- AFLW [45] (Annotated Facial Landmarks in the Wild): Real-world images including faces with variations in pose, lighting, expression, ethnicity, age, and gender (25,993 annotated faces from 21,997 images).
- 300-W [46] (300 Faces-In-The-Wild): In-the-wild images from indoor and outdoor scenes including variations in identity, expression, illumination, pose, occlusion, and face size (600 annotated faces from 399 images).

2) FACE DETECTION AND ALIGNMENT METHODS

Among state-of-the-art face detection and alignment models, the Multi-Task Cascaded Convolutional Network (MTCNN) [38] has a cascaded CNN architecture of three different networks: (1) A Proposal Network (P-Net) where several facial regions in the image are proposed as candidates; (2) A Refinement Network (R-Net) where all candidate regions are rejected or retained for further analysis by the following network; (3) An Output Network (O-Net) where remaining candidates are further refined to obtain a final selected region corresponding to the face region with landmarks. At each stage, bounding box regression vectors and non-maximum suppression are computed to obtain corresponding outputs. This model was trained on three different datasets (WIDER FACE, FDDB, and AFLW) and performs joint face detection and alignment with resulting 5-point landmarks.

As opposed to MTCNN's regression-based approach, the Convolutional Experts Constrained Local Model (CE-CLM) uses a model-based approach where the appearance of facial landmarks are computed to obtain an output [40]. Traditional CLMs use local detectors to model each facial landmark and shape them from constrained optimization techniques. Although this approach can be robust to occlusions or subject pose (especially faces in profile), it is severely impeded by complex variation in facial appearance such as facial hair, makeup, or accessories. Most of these complex variations should not occur in NICU-based data, thus warrants further exploring for a neonatal population. The Convolutional Experts Network (CEN) can model such variations using a mixture of experts.

The CE-CLM framework can be considered a three-fold process; first, a face detector is applied to obtain landmark positions (CLM); second, each landmark is accurately localized (CEN); and third, all landmarks are properly aligned using point distribution models to create a 68-point facial landmarks. CE-CLM can use different model architectures for its backbone including cascade detectors, tree-structured models, and more recently the MTCNN model. The CE-CLM model was trained on four different datasets (300-W, 300-VW, IJB-FL, and Menpo Challenge), that were selected due to the presence of challenging environment such as varying lighting, occlusions, different image quality, varying poses, profile faces, and video data [40].

Addressing the estimation of facial pose, the img2pose model [47] proposes a 6-degree-of-freedom (6DoF) model for each detected face in an input image. Compared to common face detectors, the img2pose does not rely on face bounding boxes or facial landmarks. Instead, it first aligns the 6DoF facial model to the 3D face pose and then projects the model onto the image to obtain a bounding box as a by-product. The authors propose various settings of size and shape for fitting the box around the person's face. The img2pose model leverages the Faster R-CNN detector [28] (with ResNet-18 as a backbone [48]) to propose areas in the image as candidates for face locations. From these proposed regions, features are extracted for face classification and 6DoF face pose regression. The img2pose model is trained and tested on the WIDER FACE dataset for 2D face detection evaluation.

Aiming to obtain dense face localisation, the RetinaFace model [49] is a single-stage detector that uses a multi-task network for face classification, face box regression, 5-point facial landmark regression and 1k 3D vertices regression. The RetinaFace model uses the ResNet-50 model as a backbone for generating a feature pyramid, applies a context module

to each pyramid level to increase the receptive field to help detect smaller faces, and uses different anchor sizes at each level to detect faces of varying sizes. A multi-task loss is computed as a linear combination of the loss of each corresponding task. Deng *et al.* demonstrated that each of these tasks can contribute to one another. The RetinaFace model is trained and tested on the WIDER FACE dataset for face detection evaluation, with an emphasis on the Hard subset.

Most recently, the YOLO5Face [39] has redesigned the YOLOv5 [32] object detection model into a face detector. Important modifications were implemented such as adding a 5-point landmark regression head to obtain facial alignment, reducing the kernel sizes in the spatial pyramid pooling (SPP) block to enable detection of smaller faces, replacing YOLOv5's Focus layer with a Stem block to improve generalization and reduce computational complexity, and tailoring the data augmentation techniques to face detection. Qi *et al.* have provided different YOLO5Face models based on various YOLOv5 backbones for computer or mobile device applications. The overall loss function of YOLO5Face extends from YOLOv5 as a compound loss of bounding box location regression loss, confidence loss, classification loss, plus a Wing loss for the added landmark regression. YOLO5Face used WIDER FACE to train and test the face detection task.

In this paper, we use the MTCNN, CE-CLM, img2pose, RetinaFace, and YOLO5Face pretrained face detection models. These models were selected due to their variety in architecture, different adult-based datasets used during their development, different landmark regression approaches for obtaining a sparse 5-point or dense 68-point facial landmark, and different approaches for obtaining face bounding boxes. In total, this collection represents a broad cross-section of the state of the art in face detection models.

III. DATASETS

This section describes the three neonatal datasets used in this paper. A summary of the datasets is provided in Table 1.

A. CHEO

This section describes the image data collection and preparation for face detection from video data collected at the Children's Hospital of Eastern Ontario (CHEO).

1) DATA COLLECTION

As part of an overarching non-contact neonatal monitoring research, about 153 hours of video recordings and physiologic data were collected from 33 newborns admitted at the NICU of CHEO. A depth-sensing camera, the Intel RealSense SR300, was placed above the patient to capture color, depth, and near-infrared data for up to 6 hours per patient during continuous neonatal monitoring. Only RGB data was used in this study. Each patient was recorded in one of the three different bed types: incubator, crib, and overhead warmer. Given the purely observational design of this study, video data captured challenging scenes including complex patient poses, facial occlusions from hospital equipment and free-moving limbs, diverse lighting conditions, clinical interventions, routine care procedures, and varying camera view points. This study was approved by the Research Ethics Boards of both the hospital and Carleton University (CU-117311, CU-107193). Unfortunately, we are not able to publicly release the CHEO dataset due to restrictions from the hospital's Research Ethics Board.

2) DATA EXTRACTION

One image was extracted per 30 seconds of video data. This provided substantial variation during events (*e.g.*, clinical intervention, patient motion) but insufficient variety when the patient is at rest. Therefore, images were further filtered to eliminate highly similar images.

3) IMAGE HASHING

To remove visually similar images, an average hash method was used. Each image was resized to 8×8, grayscaled, and the average of this new image is computed. Each pixel is then compared to the calculated average to compute a bit value (*e.g.*, set to 1 if above the average, and 0 otherwise) and all bits are extracted sequentially to form a 64-bit integer as the image hash. Images were then hierarchically clustered using hamming distance to compare hash values and only one image from each cluster was retained such that no two images had a hamming distance ≤ 5.

4) DATA CURATION

The CHEO image set was subdivided into "optimal", "challenging", and "negative" data subsets. The "optimal" subset (CHEO_{opt}) includes images where the patient's face is clearly visible, with high lighting, no facial occlusion, clear frontal view, close distance from the camera (max 60 cm), no ongoing phototherapy treatment, no ongoing clinical intervention, and no blur due to patient motion. The "challenging" subset (CHEO_{ch}) included the opposite cases from the "optimal" subset. The "negative" set was excluded from further analysis and contained those images where the face of the patient is not visible, such as complete facial occlusion, face out of frame, patient absent from bed, or complete darkness making it impossible to visualize the patient's face for a human observer.

5) STANDARDIZED FACE ORIENTATION

The camera is typically at a fixed position and orientation for the entire recording session, but the orientation varied between patients. As a preprocessing step, all images were rotated such that the head is at the top of the image (referred to as the "North" orientation).

6) FACE ANNOTATION

Faces within each image were manually annotated. Bounding boxes captured the area from forehead to chin and ear to

TABLE 1. Description of datasets.

Dataset	*Tot Imgs*	*Unique Imgs*	*Patients*	*Age*	*Resolution*	*Avg. BBox Area*	*Viewpoint*
COPE	288	183	27	18h - 3d	3008 x 2000	23%	Close up face
NBHR	889	565	257	0 - 6d	640 x 480	15%	Close up face
$CHEO_{optimal}$	2,048	111	16*	4 - 64d	640 x 480	8%	Full body
$CHEO_{challenging}$	11,517	1,855	33	4 - 64d	640 x 480	3%	Full body
Total	14,742	2,714	317	18h - 64d	–	–	Multiple views

*Subset of patients from the entire CHEO dataset, including only images representing optimal conditions (see text).

ear, and only visible parts were selected in cases of partial occlusions.

B. COPE

The Infant Classification of Pain Expressions (COPE) dataset was obtained from Brahnam *et al.* [50], [51] where their research focused on neonatal pain assessment. The database contains 288 images of 27 newborn faces in the NICU, collected during a painful procedure (*e.g.*, heel lancing), vs non-painful ones (*e.g.*, light puff of air on the nose, friction from rubbing alcohol).

To finalize data preparation, **standardized face orientation**, **image hashing**, and **face annotation** were performed similarly as described in Section III-A.

The face annotations created here are available at github.com/GreenCUBIC/NICUface and researchers are invited to inquire with Brahnam *et al.* [50], [51] for access to the original image dataset.

C. NBHR

The newborn baby heart rate estimation database (NBHR) was obtained from Huang *et al.* [5] where they collected synchronized video recordings and physiologic signal for non-contact neonatal heart rate estimation. The database includes 9.6 h of facial videos collected among 257 patients, with photoplethysmograph (PPG) signals, heart rate values, and oxygen saturation levels.

The dataset consisted of 1130 videos, where for each video, the first frame was extracted as an image. To finalize data preparation, **standardized face orientation**, **image hashing**, and **face annotation** were performed similarly as described in Section III-A.

The face annotations corresponding to the NBHR extracted images and the detailed image extraction protocol are available at github.com/GreenCUBIC/NICUface and researchers are invited to inquire with Huang *et al.* [5] for access to the video dataset.

IV. METHODS

This section describes the pretrained models used in this study, in addition to supplemental analysis of complex scenes (Section IV-A). Finetuned models are then created using the best pretrained networks to create our NICUface models (Section IV-B). Evaluation of all face detection models is presented (Section IV-C), followed by the description of our face orientation estimation methods.

A. DATA ANALYSIS FROM PRETRAINED MODELS

1) PRETRAINED MODELS

Five pretrained models are used here: MTCNN, CE-CLM, img2pose, RetinaFace, and YOLO5Face.

a: MTCNN

The pretrained MTCNN model is tested without modification on our neonatal datasets(see Section II-B-2 and [38] for further details).

b: CE-CLM

The CE-CLM can use different face detectors as a backbone of the CEN network to obtain landmark positions. This paper leverages the MTCNN model for the CEN backbone. Predictions differ from the MTCNN model in that they are further refined and also include the 68-point face alignment.

c: img2pose

The pretrained img2pose is tested without modification (see Section II-B-2 and [47] for further details), and using the bounding box setting encapsulating the face from forehead to chin.

d: RetinaFace

The pretrained RetinaFace was used with a ResNet-50 backbone model, as described in Section II-B-2 and [49].

e: YOLO5Face

The YOLO5Face model was used with the "large" Stem block since this was shown to be one of the most accurate by Qi *et al.* [39]. The YOLOv5l6 is not used here since they reported that, while the P6 block addition improved performance on the WIDER FACE's Easy and Medium subsets, it can decrease the performance on the Hard subset (which more closely resembles our data).

The five pretrained models were tested in MATLAB using an NVIDIA GeForce GTX 1070 GPU, and with Python using a Tesla P100-PCIE-16GB.

2) COMPLEX NICU SCENES

Complex scenes are further analyzed by evaluating face detection performance under various clinical challenges.

TABLE 2. Train & test sets.

Split	Train	Test
1	COPE + NBHR	$CHEO_{opt}$
2	COPE + $CHEO_{opt}$ + $CHEO_{ch}$	NBHR
3	NBHR + $CHEO_{opt}$ + $CHEO_{ch}$	COPE
4	COPE + NBHR + $CHEO_{opt}$ + $CHEO_{ch\_\notin F1}$	$CHEO_{ch\_F1}$
5	COPE + NBHR + $CHEO_{opt}$ + $CHEO_{ch\_\notin F2}$	$CHEO_{ch\_F2}$
6	COPE + NBHR + $CHEO_{opt}$ + $CHEO_{ch\_\notin F3}$	$CHEO_{ch\_F3}$
avg(4:6)	—	$CHEO_{ch}$

Using the $CHEO_{ch}$ dataset, we extract challenging cases based on varying levels of occlusions, viewpoints, and lighting. In terms of occlusions, they can occur when the patient is sucking on a soother, from the nurse's hand or arm during a clinical intervention, when the patient is wearing a phototherapy eye mask during treatment, from a ventilation support device, or from free-moving limbs or beddings. Viewpoints are considered to be challenging when the camera is positioned at a far distance from the patient (>1m); when the patient is being held in the bed; or when the face is only visible in profile view, from a near-top view, or from near-back view when the patient is in prone position. Lighting conditions are challenging during dimly lit periods (*e.g.*, patient sleeping or reduced sensory input environments) or during phototherapy treatment. To evaluate these complex scenes, the best performing pretrained models (RetinaFace and YOLO5Face) are tested on each challenging case before being finetuned. Previous neonatal monitoring applications have discussed how challenges from clinical scenes can pose a problem to the face detection performance. This paper quantifies the impact of these individual complex scenes.

B. FINETUNED MODELS

We create the NICUface detectors from finetuning pretrained RetinaFace and YOLO5Face models. For both models, models were trained and evaluated on different patient subsets to quantify model generalization, as described in Table 2. Note that the same 16 patients from $CHEO_{opt}$ were present in $CHEO_{ch}$, only with different challenging scenes. Split 1 therefore only trained on COPE + NBHR and tested on $CHEO_{opt}$ to maintain testing this dataset with entirely different patients that were not seen during training. For the $CHEO_{ch}$ data, given its variety of challenging conditions, the 17 unique patients in this dataset (not present in $CHEO_{opt}$) were divided into three folds. Each fold contained a proportional amount of complex scenes, especially considering low lighting and patients on ventilation support. The final reported performance on the $CHEO_{ch}$ dataset is reported as the average of these three folds for the pretrained and finetuned models to provide a fair comparison.

1) NICUface-RF

The RetinaFace model was finetuned using a ResNet50 backbone and the RetinaFace weights. Finetuning occurred over 10 epochs with a batch size of 8 with an initial learning rate of 0.001 with a warmup to 0.1 at epoch 1 and then 0.1 decay for epochs 2 and 5. Anchors were matched to an object when the intersection over union (IOU) was larger than 0.45 and to the background when the IOU was less than 0.3. Training data were augmented with random horizontal flip and photo-metric colour distortion. The loss function was not changed from the original RetinaFace model; however, during training landmark regression error was ignored by setting all landmark inputs in the training data to -1. Only bounding box error was used.

2) NICUface-Y5F

The YOLO5Face model was finetuned using the YOLOv5l weights, trained over 10 epochs with batch size of 16, initial learning rate of 0.0032 and final learning of 0.12, optimized using stochastic gradient descent with 0.5 momentum in the first 2 epochs and momentum of 0.843 after, and an IOU threshold of 0.2 during training. The loss function of NICUface-Y5F is similar to the loss function of YOLO5Face as,

$$loss = loss_{box} + loss_{conf} + loss_{cls} + \lambda_{land} \cdot loss_{land} \quad (1)$$

where $loss_{box}$ is the bounding box regression loss, $loss_{conf}$ is the confidence loss, $loss_{cls}$ is the classification loss, and $loss_{land}$ is the landmark regression loss with weighting factor λ_{land}. This λ_{land} was set to only 0.005 to pay less attention to the landmarks given the unsupervised landmark localisation. Similarly to YOLO5Face, the $loss_{conf}$ and $loss_{cls}$ were optimized using the cross-entropy loss function. In terms of data augmentation, YOLO5Face reported that Mosaic augmentation and removal of up-down flipping improved their performance, but only on the Hard WIDER FACE subset without ignoring small faces or random cropping. Given the difficulty of our neonatal dataset we also applied Mosaic and removed up-down flipping.

The training set was divided into two sets of data used during training and validation stages where different patients were used for training and validation. As can be seen in Table 2, each face detector was tested on a completely different dataset from that used to train the models.

C. FACE DETECTION EVALUATION

For face detection performance of pretrained and finetuned models, all models are evaluated using the average precision metrics with varying intersection over union (IOU) requirements. The AP is calculated with IOU $\geq$ 0.5 as a standard evaluation metric (AP50), while the mAP captures the mean over AP with IOU=0.5:0.05:0.95 to reward models producing more specific bounding boxes. The facial landmarks are not evaluated quantitatively here since no gold standard landmark annotations were performed on our neonatal datasets; however, landmarks are reviewed qualitatively to generally assess the performance of the face alignment task and to identify challenging cases where the alignment would fail.

In evaluating all models, we opted to only output the prediction with the highest confidence score. This approach is feasible since only one face is assumed to be present in each image. Given the difficult task of finding neonatal faces in complex scenes, this approach allows low confidence predictions of the patient's face to still be considered while ignoring other irrelevant false predictions in the scene.

We also look at cases where we decrease the IOU threshold to 0.3 (AP30) to include slightly overestimated or underestimated bounding boxes around the face. Although most object detectors report AP with IOU of at least 0.5, recent applications leveraging these detectors have opted for lower IOU threshold in cases where the objects are small and hence the AP/mAP metric would be drastically impacted by marginal errors [52], [53].

D. FACE ORIENTATION ESTIMATION

In many neonatal monitoring applications, a preprocessing step is required where images are rotated to standardize the orientation of the face. Having the patient's face oriented North facilitates the face detection and alignment task, and this rotation step is often performed manually (laborious) or programmatically through trial and error by rotating the image at 90° increments until a face is detected (unreliable if a face is detected at non-North direction without providing confidence from this orientation). We therefore propose a face orientation estimation approach where the image is rotated in four 90° increments. The ''North'' face orientation is predicted as the direction that produces the most confident bounding box, the most coherent facial landmark positions, or both:

1) FACE ORIENTATION ESTIMATION BASED ON FACE BOX CONFIDENCE SCORE

From the detected bounding boxes in all four directions, we predict that the North-facing orientation should have the highest confidence score.

2) FACE ORIENTATION ESTIMATION BASED ON FACIAL LANDMARK POSITION

From the detected 5-point facial landmarks in all four directions, we predict that the North-facing orientation should have landmarks positioned in manner where the nose is below the eyes. We measure the Nose-to-Eye-Line Angle (NELA) which measures the angle of a line that originates at the nose landmark and intersects the inter-ocular line at 90°. A North-facing orientation would result in a NELA of 90° (± 45° to account for minor pose variations). This method is illustrated in Fig. 3.

3) FACE ORIENTATION ESTIMATION BASED ON COMPLETE FACIAL DETECTION

Given the strength and weaknesses of each technique presented above, a final and more comprehensive face orientation estimation approach is presented by leveraging both the face box confidence scores and the facial landmark positions. The North orientation is determined by selecting the detection with the highest confidence score that also has a valid NELA.

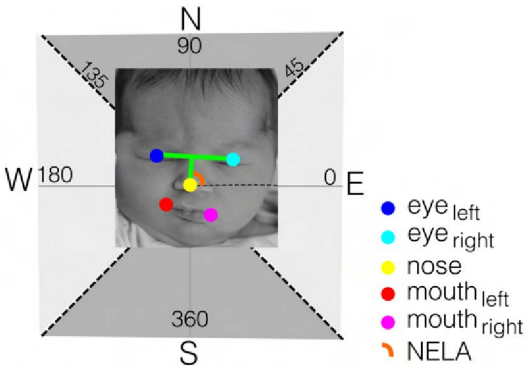

FIGURE 3. Face orientation estimation from landmark position.

V. RESULTS AND DISCUSSION

This section assesses the state of the art in face detection for neonatal patients in NICU environments. Experiments cover neonatal face detection challenges from multiple experiments with different pretrained models, datasets, and complex NICU scenes (Section V-A to V-C). Face orientation estimation is then evaluated (Section V-D).

A. PRETRAINED MODELS AND NEONATAL DATASETS

Among all pretrained models presented in the top half of Table 3, MTCNN performs worst, and interestingly, the CE-CLM model using MTCNN as a backbone detector performs better in comparison. The fact that landmark positions are refined in the CE-CLM model before applying the denser 68-point distribution model strongly suggests the advantage of the CEN layers in the localisation task. Figure 4 depicts results from all models with increasing level of scene complexity from left to right, and increasing performance of each model from top to bottom. Bounding box predictions are labelled as correct (IOU≥50, Green), partial (IOU≥30, Yellow), or incorrect (IOU<30, Red). As illustrated in Fig. 4, some false negatives with MTCNN have become true positives with CE-CLM for the COPE dataset (with correct detection and decent facial alignment despite the partial occlusion by blanket), for the NBHR dataset (with partial detection and misaligned facial landmarks due to profile view), and for the CHEO_{opt} dataset (with partial detection and proper facial alignment). For the CHEO_{ch} dataset, no detection is obtained with MTCNN and CE-CLM for most scenes, except for a few with very minor occlusions and viewpoints where all facial landmarks are visible (e.g., patient imaged from a far distance).

While CE-CLM revealed improved results compared to MTCNN, the value of generating 68-point landmarks is likely to be application-dependent. For example, such fine-detailed facial structure is not needed for HR estimation or jaundice detection (which primarily look at the skin),

TABLE 3. Face detection results.

	COPE			NBHR			$CHEO_{optimal}$			$CHEO_{challenging}$		
Model	AP30	AP50	mAP	AP30	AP50	mAP	AP30	AP50	mAP	AP30	AP50	mAP
MTCNN	74.31	74.31	51.79	60.07	59.74	38.18	49.95	48.83	26.41	7.32	4.93	1.62
CE-CLM	92.08	91.19	31.94	79.34	77.29	25.81	64.10	57.77	16.35	16.19	8.95	1.75
img2pose	88.85	94.36	65.90	87.18	92.37	56.42	87.21	94.49	65.14	49.13	50.46	26.76
RetinaFace	100	100	76.73	100	100	78.60	99.95	99.95	79.12	52.47	52.12	29.56
YOLO5Face	100	100	83.80	99.56	99.67	77.95	95.73	95.73	76.39	50.78	48.58	28.65
NICUface-RF	100	100	86.55	100	100	80.53	99.10	99.10	77.92	86.12	79.73	43.67
NICUface-Y5F	100	100	88.30	99.95	99.95	82.39	93.16	93.16	76.73	88.61	87.98	61.75

but it would be highly relevant for pain assessment or sleep-wake detection (which primarily look at the facial expression). This could open a door to retraining an 5- or 68-point landmark distribution model suitable for the neonatal population with 5 or 68 salient facial features observed in newborns, respectively.

Overall, the pretrained RetinaFace and YOLO5Face methods outperform all other approaches, with consistent detection (near 100% in AP30 and AP50) for the COPE, NBHR, and $CHEO_{opt}$ datasets. For these three datasets, all detections are correct, with proper facial alignment despite the minor occlusions or the patient being viewed in profile. In a complementary manner, RetinaFace performs best on NBHR, $CHEO_{opt}$, and $CHEO_{ch}$, while YOLO5Face performs best on COPE, as demonstrated in Table 3.

Given that the pretrained RetinaFace and YOLO5Face models consistently outperformed the MTCNN and CE-CLM methods across all datasets, the MTCNN and CE-CLM methods were not investigated further. Similarly, the img2pose results are significantly worse than RetinaFace and YOLO5Face on COPE, NBHR, and $CHEO_{opt}$ datasets. However, at first glance, results for img2pose on the difficult $CHEO_{ch}$ dataset appear to be on par with RetinaFace, and YOLO5Face. Performance of these three methods across each individual complex scene are investigated in detail in the following section, before implementing the ultimate solution: NICUface.

Across all models, a consistent pattern exists in dataset performance with COPE > NBHR > $CHEO_{opt}$ > $CHEO_{ch}$. This pattern agrees with a qualitative assessment of the level of difficulty among our datasets in analogous fashion to the WIDER FACE dataset's easy, medium, and hard subsets [43]. Our COPE data represent our ''easy'' subset with close up facial views, NBHR has ''medium'' difficulty with close up faces and more challenging poses and occlusions, $CHEO_{opt}$ is ''medium-hard'' where the image includes the full body and bed environment. Finally, $CHEO_{ch}$ is a ''hard'' dataset as it includes the entire bed environment and complex scenes, such as low lighting, ventilation support, pose variation, etc. Considering that the performance of all models is significantly reduced on the $CHEO_{ch}$ dataset, the complex scenes therein are further analyzed in Section V-B. The datasets are then leveraged for the implementation of NICUface using the best competing pretrained models in Section V-C.

B. COMPLEX NICU SCENES

Among all datasets, $CHEO_{ch}$ had the lowest face detection performance for all models due to the complexity of scenes included therein. The increasing level of difficulty among these complex scenes is presented in Table 4 and Fig. 1 with varying levels of occlusions, viewpoints, and lighting. Both RetinaFace and YOLO5Face demonstrated a similar pattern of performance across the complex scenes. Mouth occlusions from a soother are not as challenging as partial occlusions from the nurse's arm/hand or from beddings. Near complete facial occlusions from the ventilation support remain the most challenging occlusion-based scenes. Among viewpoints, far distance and profile view performed best, but near-top view or prone position are most challenging given that only a small portion of the face is visible. From lighting environment, a low lighting environment doesn't affect the model as severely as the phototherapy light. Note that the phototherapy eye mask is also an occlusion-based challenge; however, we dimmed the blue-colored lighting more important to this unique scenario.

Interestingly, even though img2pose performed similary to RetinaFace and YOLO5Face on the overall $CHEO_{ch}$ dataset, closer inspection of the performance of each model across each of the NICU-specific challenges (see Table 4) reveals that img2pose is only largely outperforming the other two methods on the ''near top view'' scenario; in all other cases, img2pose under-performs. For this reason, combined with its inferior performance on the easier datasets (COPE, NBHR, and $CHEO_{opt}$), img2pose was not considered further in this study.

Having established the limits of the state of the art in face detection for complex NICU scenes, we turn our attention to addressing the remaining NICU-specific challenges through

FIGURE 4. Face Detection Results from all neonatal datasets, pretrained models, and NICUface models.Increasing level of scene complexity is demonstrated from left to right. Increasing performance of each model is presented from top to bottom. Predictions are labelled as correct (IOU≥50, Green), partial (IOU≥30, Yellow), or incorrect (IOU<30, Red).

TABLE 4. Face detection from complex NICU scenes (AP30 on $CHEO_{ch}$ with RetinaFace & YOLO5Face).

Challenge	img2pose	RetinaFace	YOLO5Face	#Imgs
OCCLUSIONS				
soother	83.72	99.50	96.58	24
intervention	52.99	60.96	64.48	88
bedding/self	50.18	61.29	53.95	286
ventilator	14.18	20.33	3.04	195
VIEWPOINT				
far distance	59.35	67.90	84.17	134
profile	56.95	68.83	69.96	47
near top view	83.32	40.45	42.24	24
prone position	0	7.34	1.11	18
LIGHTING				
low lighting	37.01	75.93	41.39	36
phototherapy	41.93	33.93	33.33	19

finetuning the most promising pretrained models (RetinaFace and YOLO5Face), leading to the NICUface models.

C. NICUface

In this section, we report on the performance of the NICUface models, where we have finetuned the top-performing RetinaFace and YOLO5Face models for NICU-specific challenges (see Table 2 for datasets used for finetuning and evaluation). From the results of the pretrained models, it was established that the RetinaFace and YOLO5Face already performed very well across COPE, NBHR, and $CHEO_{opt}$ datasets. Given the near-perfect performance of these models across these datasets, it is unsurprising that comparable results were obtained with the NICUface models, with near 100% AP30 and AP50 values among these three datasets. The advantage of fine-tuning becomes apparent on the $CHEO_{ch}$ dataset, where the NICUface models demonstrated large improvements on this challenging data. NICUface-RF showed an increase of +33.65, +30.67, and +17.74 in AP30, AP50, and mAP respectively compared to RetinaFace. NICUface-Y5F showed an increase of +37.83, +39.40, and +33.10 in AP30, AP50, and mAP respectively compared to YOLO5Face. Between both NICUface models, NICUface-Y5F slightly outperformed NICU-RF on $CHEO_{ch}$ with a difference of +2.49, +8.25, and +18.08 in AP30, AP50, and mAP, respectively.

As illustrated in Fig. 4 and in Table 5, the NICUface models showed robustness to the presence of ventilation support and patients in near-back view when in prone position, while pretrained models were impaired by these scenes. Given that

TABLE 5. Face Detection from Complex NICU Scenes (AP30 on $CHEO_{ch}$ with NICUface-RF and NICUface-Y5F).

Challenge	NICUface-RF	Challenge	NICUface-Y5F
profile	**100**	profile	**100**
soother	**100**	soother	99.23
near top view	**100**	near top view	98.24
bedding/self	**98.79**	bedding/self	97.57
low lighting	**89.61**	far distance	**97.02**
far distance	89.01	intervention	**95.21**
intervention	87.92	low lighting	87.42
prone position	76.08	prone position	**79.42**
ventilator	55.80	ventilator	**65.87**
phototherapy	**0**	phototherapy	**0**

these two complex scenes are the two most challenging ones, NICUface-RF demonstrates impressive performance with an improvement in AP30 of +68.74 and +35.47 for the prone position and ventilation support, respectively. NICUface-Y5F also improves drastically with AP30 of +78.31 and +62.83 for the prone position and ventilation support, respectively.

Moreover, both models are highly complementary to one another. NICUface-RF presents strengths in detecting patients in low lighting conditions (with +13.68 improvement in AP30), while NICUface-Y5F is better at detecting smaller faces (with +12.85 improvement in AP30). These conditions are illustrated in Fig. 4, where NICUface-RF was able to rectify RetinaFace's false positive by correctly detecting the face of the patient under very low lighting. In the same scenario, YOLO5Face also made an incorrect prediction but NICUface-Y5F was not able to rectify this error. On the other hand, during a clinical intervention, the face of the patient captured from a far distance was detected with NICUface-Y5F, while NICUface-RF avoided a previous false positive but overestimated the bounding box area. This improvement is still remarkable, given that it was now able to make a detection in the general location of the face, however it fails to reach the precision of NICUface-Y5F. Future work could investigate this complementary pairing through an ensemble network combining both models' strengths into one.

Among all evaluation metrics, the AP30 is the most reliable measure of model performance for neonatal monitoring applications. In our case, the frequent presence of small faces in the CHEO dataset warrants evaluating with a smaller threshold than the standard AP50. As seen in Table 1, our most challenging dataset has an average bounding box area that only makes up 3% of the image. In such cases, NICUface would tend to slightly overestimate the bounding box area which would severely affect the AP metric, despite the relevant prediction. Slight overestimation is not an issue for monitoring applications requiring the entire face for facial expression analysis in pain assessment, sleep-wake cycle detection, or face recognition. Slight underestimation is also not an issue when small facial ROI can be sufficient in some applications such as HR estimation or jaundice detection relying on visible skin patches. Due to high level of facial occlusions in the NICU, some non-contact neonatal monitoring applications have opted for different techniques to only obtain visible facial area (*e.g.*, skin segmentation). The AP30 metric is therefore a most reliable measure since lowering the IOU threshold permits considering slightly overestimated or underestimated predictions which can still be useful in a wide array of neonatal applications.

Note that for RetinaFace and YOLO5Face, lightweight models suitable for detections on embedded or mobile devices were also implemented using MobileNet-0.25 and ShuffleNet backbone models, respectively. These pretrained models were not investigated here since in our application, we are not limited in compute power, so these lightweight models are not particularly useful. Future work could however use these models for the implementation of other neonatal monitoring applications (*e.g.*, in home monitoring or in intelligent monitoring applications from smartphones).

Training the NICUface models took approximately 1 hour for each of the six cross validation sets (sets are listed in Table 2). However, since training can typically be done offline, when considering methods for real-time deployment our biggest consideration is the inference time required to process a single image. For the NICUface models, inference time is currently ∼2 s per image. However, it is expected that this time could be further reduced through careful optimization, the use of low-cost face tracking with periodic *de novo* detections, or the use of more powerful dedicated hardware should more frequent face detections be required.

Among all neonatal monitoring applications presented in Section I, Awais *et al.* [14] represents the only study performing automatic face detection and reporting its performance (to the best of our knowledge). They achieved 98.5% accuracy using the Fluke TiX580 camera for intensity-based face detection on patients with 0 degree head tilt (i.e., frontal view). In comparison, NICUface-RF and NICUface-Y5F achieve 100% on our COPE dataset which most closely compared to their dataset. For more challenging scenes, NICUface-RF still performs remarkably well with 100%, 99.1%, and 73.87% for NBHR, $CHEO_{opt}$, and $CHEO_{ch}$, respectively. NICUface-Y5F also performs well with our most challenging data with 100%, 88.29%, and 83.77% for NBHR, $CHEO_{opt}$, and $CHEO_{ch}$, respectively.

For both NICUface models, the blue-colored light during phototherapy treatment (in addition to the facial occlusion from the eye mask) posed a challenge for face detection. Interestingly, their pretrained counterparts were able to detect a few images when the nose and face were visible, resulting in an AP30 of ∼ 33% for both. NICUface-Y5F shows promise with very small detections from the visible skin in a few images, however with an IOU < 0.3. To address this challenge, a pre-processing technique is proposed to reduce the blue hue. The detection of ongoing phototherapy treatment (compared to patients under natural lighting) is a problem previously solved by Souley Dosso *et al.* [54], and

we leverage that work to address face detection during this complex scene in the following section.

1) FACE DETECTION ON PHOTOTHERAPY PATIENTS

This proposed technique to address face detection on phototherapy patients can be performed in three simple steps:

1) Detect phototherapy images
2) Apply blue filtering for phototherapy images
3) Face detection using NICUface

a: PHOTOTHERAPY DETECTION

The phototherapy classification presented in [54] is leveraged here to differentiate phototherapy images from those captured in natural lighting during inference.

b: BLUE FILTERING

The phototherapy classification in [54] demonstrated how the Red, Green, and Blue channels in the natural images are almost uniformly distributed. In comparison, phototherapy images are heavily weighted with blue-colored pixels, relative to red-colored pixels. This important knowledge is exploited here to perform a color space transformation on the phototherapy images to equalize the colour channels. Our ''Blue Filtering'' method scales pixel intensities of the red and blue channels to match the pixel intensities of the green channel to equalize the image as

$$Scale_R = \frac{1}{w \times h}\left[\sum_{i=1}^{w}\sum_{j=1}^{h} G_{ij} - \sum_{i=1}^{w}\sum_{j=1}^{h} R_{ij}\right] \quad (2)$$

$$Scale_B = \frac{1}{w \times h}\left[\sum_{i=1}^{w}\sum_{j=1}^{h} B_{ij} - \sum_{i=1}^{w}\sum_{j=1}^{h} G_{ij}\right] \quad (3)$$

$$R^*_{ij} = \min\left(255, R_{ij} + Scale_R\right) \quad (4)$$

$$B^*_{ij} = \max\left(0, B_{ij} - Scale_B\right), \quad (5)$$

where $Scale_R$ measures the scaling factor using G_{ij} and R_{ij}, the Green and Red channels, respectively, for pixels at the i^{th} position among image width (w), and j^{th} position among the image height (h). R^*_{ij} represents the updated Red channel in the ''equalized-phototherapy'' image, as illustrated in Fig. 5. Given that the $Scale_R$ value is added to each pixels, the updated R^*_{ij} caps all pixels exceeding 255 as an intensity of 255. Similar steps are performed to obtain an updated Blue channel B^*_{ij} by scaling down the pixels by $Scale_B$, and capping all pixels inferior to 0 as an intensity of 0. Note that this Blue Filtering approach follows the intuition from [54] where the Blue channel is over-expressed and the Red channel is under-expressed, thereby scaling them to match the Green channel aiming to equalize the image. Scaling the Red and Green channel to match the Blue channel would produce a similar outcome (as well as scaling the Blue and Green channel to match the Red channel).

As we can see in Fig. 5, the predominant blue hue is successfully reduced, especially in areas of visible skin. Other surfaces still have a slight blue tint; this is apparent in areas known to be truly white, such as the bedding or eye mask.

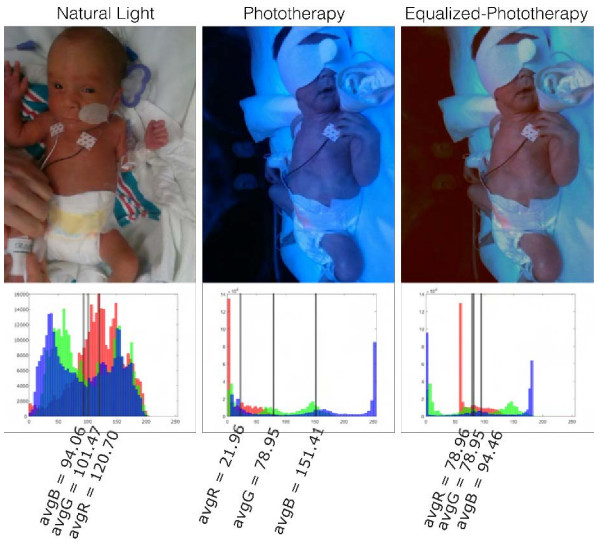

FIGURE 5. Blue filtering of phototherapy images. The average of each corresponding RGB channel are sparse in the phototherapy image, and thereby attempts to narrow the gaps across channels in the equalized-phototherapy image simulating the natural light condition.

c: FACE DETECTION WITH NICUface AND BLUE FILTERING

During inference, the phototherapy detection is applied directly on the image to differentiate between lighting environments. Images deemed to represent ongoing phototherapy are processed using the Blue Filtering method and NICUface detects the face in the modified image. Images deemed to reflect natural lighting are unchanged. To validate this face detection approach on phototherapy patients, the 19 images from the only patient in our $CHEO_{ch}$ dataset are used and evaluated with AP30 metric. This method is validated using the best performing face detectors (RetinaFace, YOLO5Face, NICUface-RF, and NICUface-Y5F).

Face detection results are demonstrated in Table 6. For the state-of-the-art models, results in AP30 improved by +2.47 and +16.22 for RetinaFace and YOLO5Face, respectively, with Blue Filtering. The benefit of Blue Filtering is more apparent with the NICUface models where AP30 is increased by +50.00 and +41.49 for NICUface-RF and NICUface-Y5F, respectively.

These results show great promise in the use of pre-processing methodologies for color-based challenges in other machine vision applications, without requiring retraining an entire model. This is particularly useful in cases when obtaining new data can be difficult or expensive.

D. FACE ORIENTATION ESTIMATION

Since the face detection algorithms performed best on the COPE dataset, we use it with one of the best performing pretrained models, YOLO5Face, to evaluate our face orientation estimation approach. The COPE dataset and its

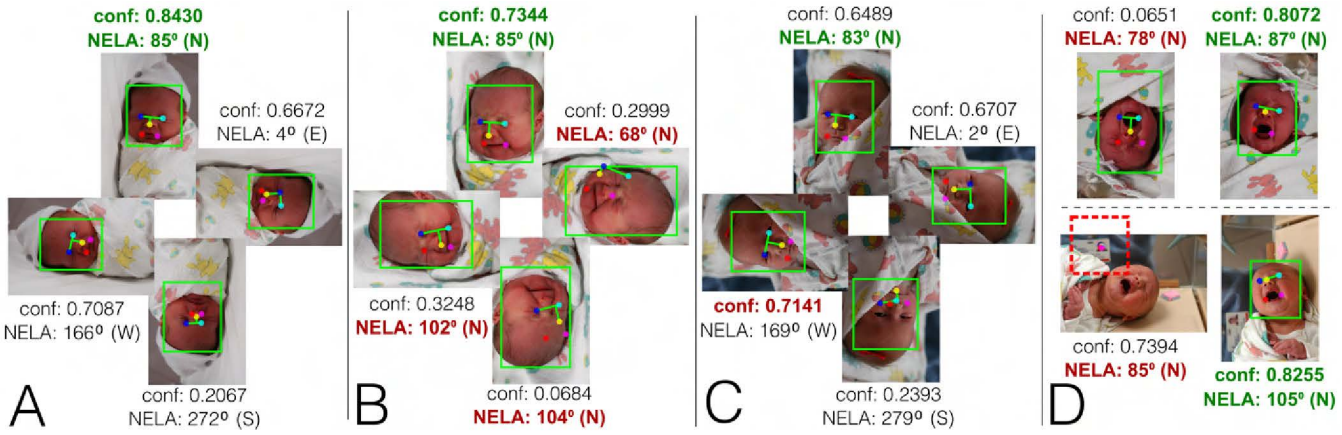

FIGURE 6. Sample face orientation predictions with face detection confidence score (*conf*) and NELA.

TABLE 6. AP30 face detection results on phototherapy patients.

Model	Without BlueFiltering	With BlueFiltering
RetinaFace	33.93	36.40
NICUface-RF	0	**50.00**
YOLO5Face	33.33	49.55
NICUface-Y5F	0	41.49

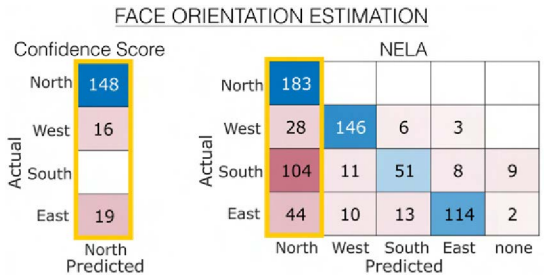

FIGURE 7. Face Orientation Predictions from YOLO5Face Confidence Scores and NELA with COPE dataset

TABLE 7. Face Orientation Estimation (COPE + YOLO5Face) using face detection confidence score (conf), NELA, or both.

Orientation	AP30	Conf	NELA	Conf+NELA
North	**100**	**80.88**	50.97	**99.45**
West	96.09	–	87.43	–
South	93.08	–	72.86	–
East	92.67	–	**91.20**	–

annotations are artificially rotated at 90°, 180°, and 270° to create sets of images with the face oriented North, West, South, and East. As observed in Table 7 and Fig. 5-D, the face orientation estimation approach based on the confidence scores alone predicts ''North'' as the North-facing face orientation 80.88% of the time, and ''West'' or ''East'' otherwise. It never predicts ''South''. The face orientation estimation based solely on landmark positions performed less accurately (50.97% precision for ''North'') since the South orientation is heavily misclassified as North. Compared to the confidence score based approach, this NELA method can detect other orientations (West, South, East) based on the NELA (at 180°, 270°, 360°/0°, respectively). However, if no face is detected, it cannot make a prediction (predicts *none*), while the confidence score approach is unaffected by this limitation.

The fused approach leverages strengths from both face confidence scores and NELA, and outperforms the individual approaches with 99.45% precision.

In ideal cases depicted in Fig. 6-A, the score-based and NELA-based methods are both effective. With more variations in facial expression, the NELA-based technique can be affected by the localisation of each landmark, as demonstrated in 6-B. In many cases, it forces the position of the landmarks to face North, no matter the actual orientation. In other cases, it predicts ''West'' or ''East'' appropriately, but the ''South'' orientation is often predicted to be ''North''. With patient occlusions, the score-based technique can be affected by a reduced face detected confidence, as demonstrated by 6-C. The combination of both confidence and NELA led to the best performance.

It is important to note that even if the detector finds a face in all orientations, facial alignment might be unreliable, and therefore not suitable for facial expression analyses. The top panel of Fig. 6-D demonstrates this, where the highest confidence score is properly detected from the North orientation compared to South. As for the landmarks, the North orientation produces a correct NELA at 87°, while the South produces an incorrect NELA at 78° given that the location of eyes- and mouth-landmarks are incorrect. Our proposed face orientation estimation approach is simple but reliable when assuming that only one face is present in the image. Additional faces might affect the desired patient's detection confidence score, as seen in the bottom panel of Fig. 6-D, where the detector found the face of the clinical staff from their ID. Despite this interesting detection, the patient's North-facing orientation still produced the highest scoring confidence and our proposed face orientation estimation method remains robust to the incorrect ID detection.

Standardizing the face orientation is an important preprocessing step in neonatal monitoring applications since it is often performed manually or as a trial-an-error approach. It is important to note that the South-facing orientation might be irrelevant for most adult-based detectors leveraging mostly upright standing or lying adults; however, in neonatal monitoring this direction is important since the patient could be repositioned in the bed, especially in cribs. The face orientation estimation model proposed here could therefore be of value to these studies.

VI. CONCLUSION

This paper evaluated the state of the art in (adult-trained) face detection models for complex NICU patient scenes. While these models leveraged challenges from an adult population including facial hair, makeup, and accessories, our neonatal population present entirely different and unique challenges such as phototherapy light, hospital equipment, clinical intervention with nurse's hands holding the face, and soother usage.

MTCNN, CE-CLM (with MTCNN backbone), img2pose, RetinaFace, and YOLO5Face performed adequately for simple scenes, where the patient face was clearly visible, in bright light, forward facing, unoccluded, and of reasonable size proportional to the image. However, these methods failed to robustly identify patient faces in complex scenes involving phototherapy lighting, ventilation support, near top view when held in the bed, and near back view when in prone position.

This study addressed these important shortcomings with the NICUface models by finetuning highly performant RetinaFace and YOLO5Face pretrained models. Our proposed NICUface models outperform previous state-of-the-art models for neonatal face detection and are robust to many identified complex NICU scenes. The most challenging scenes (prone position and ventilation support) showed exceptional improvement, demonstrating the effectiveness of finetuning state-of-the-art face detectors for our neonatal population. A solution for addressing the blue hue images from patients undergoing phototherapy treatment was also effective for detecting neonatal faces from this complex scene. On our most challenging dataset, both NICUface models are highly complementary where NICUface-Y5F works best on smaller faces and NICUface-RF on lighting environment. This paper therefore strongly suggests leveraging both NICUface models in neonatal monitoring applications for various goals.

All gold standard face annotation data, finetuned NICUface models, and face orientation estimation method are provided here at github.com/GreenCUBIC/NICUface. It is hoped that the annotation data may be used by other groups to continue to advance the state of the art in neonatal face detection, while the finetuned NICUface models and face orientation estimation will be useful to groups requiring face ROI for a variety of non-contact neonatal patient monitoring applications.

ACKNOWLEDGMENT

The authors would like to thank Joshua Tanner and students from SYSC4415 (Intro to Machine Learning) for supplying portions of the annotations for the CHEO and NBHR datasets, respectively.

REFERENCES

[1] S. Fernando, W. Wang, I. Kirenko, G. de Haan, S. B. Oetomo, H. Corporaal, and J. van Dalfsen, "Feasibility of contactless pulse rate monitoring of neonates using Google glass," in *Proc. 5th EAI Int. Conf. Wireless Mobile Commun. Healthcare*, 2015, pp. 198–201.

[2] J. H. Klaessens, M. Van Den Born, A. van der Veen, J. S.-van de Kraats, F. A. van den Dungen, and R. M. Verdaasdonk, "Development of a baby friendly non-contact method for measuring vital signs: First results of clinical measurements in an open incubator at a neonatal intensive care unit," in *Proc. SPIE*, vol. 8935, Feb. 2014, Art. no. 89351P.

[3] M. Villarroel *et al.*, "Continuous non-contact vital sign monitoring in neonatal intensive care unit," *Healthcare Technol. Lett.*, vol. 1, no. 3, pp. 87–91, Sep. 2014.

[4] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–18, Dec. 2019.

[5] B. Huang, W. Chen, C.-L. Lin, C.-F. Juang, Y. Xing, Y. Wang, and J. Wang, "A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks," *Eng. Appl. Artif. Intell.*, vol. 106, Nov. 2021, Art. no. 104447.

[6] D. G. Kyrollos, J. B. Tanner, K. Greenwood, J. Harrold, and J. R. Green, "Noncontact neonatal respiration rate estimation using machine vision," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Aug. 2021, pp. 1–6.

[7] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, "Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment," *Comput. Biol. Med.*, vol. 129, Feb. 2021, Art. no. 104150.

[8] S. Brahnam, C.-F. Chuang, R. S. Sexton, and F. Y. Shih, "Machine assessment of neonatal facial expressions of acute pain," *Decis. Support Syst.*, vol. 43, no. 4, pp. 1242–1254, Aug. 2007.

[9] S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, and T. Barrier, "Neonatal pain detection in videos using the icopevid dataset and an ensemble of descriptors extracted from Gaussian of local descriptors," *Appl. Comput. Informat.*, Jul. 2020, doi: 10.1016/j.aci.2019.05.003.

[10] M. N. Mansor, S. Yaacob, M. Hariharan, S. N. Basah, S. H. F. S. A. Jamil, M. L. M. Khidir, M. N. Rejab, K. K. M. Y. Ibrahim, A. H. F. S. A. Jamil, A. K. Junoh, and S. A. Saad, "Jaundice in newborn monitoring using color detection method," *Proc. Eng.*, vol. 29, pp. 1631–1635, Jan. 2012.

[11] W. Hashim, A. Al-Naji, I. A. Al-Rayahi, and M. Oudah, "Computer vision for jaundice detection in neonates using graphic user interface," in *Proc. IOP Conf., Mater. Sci. Eng.*, vol. 1105, no. 1, 2021, Art. no. 012076.

[12] Y. Mukai, K. Morita, N. C. Shirai, T. Wakabayashi, H. Shinkoda, A. Matsumoto, Y. Noguchi, and M. Shiramizu, "Automatic classification of neonatal sleep-wake states based on facial video analysis," in *Proc. 4th Int. Conf. Inf. Technol. Res. (ICITR)*, Dec. 2019, pp. 1–6.

[13] M. Awais, X. Long, B. Yin, S. F. Abbasi, S. Akbarzadeh, C. Lu, X. Wang, L. Wang, J. Zhang, J. Dudink, and W. Chen, "A hybrid DCNN-SVM model for classifying neonatal sleep and wake states based on facial expressions in video," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1441–1449, May 2021.

[14] M. Awais *et al.*, "Novel framework: Face feature selection algorithm for neonatal facial and related attributes recognition," *IEEE Access*, vol. 8, pp. 59100–59113, 2020.

[15] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh, "Domain specific learning for newborn face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1630–1641, Jul. 2016.

[16] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, 2012.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[18] Y. S. Dosso, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Neonatal face tracking for non-contact continuous patient monitoring," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2020, pp. 1–6.
[19] F.-T.-Z. Khanam, A. G. Perera, A. Al-Naji, K. Gibson, and J. Chahl, "Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks," *J. Imag.*, vol. 7, no. 8, p. 122, Jul. 2021.
[20] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[21] Q. Lin, R. He, and P. Jiang, "Feature guided CNN for baby's facial expression recognition," *Complexity*, vol. 2020, pp. 1–10, Nov. 2020.
[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[23] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.
[24] T. Khan, "An intelligent baby monitor with automatic sleeping posture detection and notification," *AI*, vol. 2, no. 2, pp. 290–306, Jun. 2021.
[25] B. Amos *et al.*, "OpenFace: A general-purpose face recognition library with mobile applications," *CMU School Comput. Sci.*, vol. 6, no. 2, pp. 1–20, 2016.
[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2014, pp. 580–587.
[27] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.
[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 779–788.
[30] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.
[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[32] G. Jocher *et al.*, "Ultralytics/yolov5: V6.0-YOLOv5n 'Nano' models, roboflow integration, tensorflow export, opencv DNN support," Zenodo, Cham, Switzerland, Oct. 2021, doi: 10.5281/zenodo.5563715.
[33] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10781–10790.
[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2019, pp. 6105–6114.
[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.
[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[39] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," 2021, *arXiv:2105.12931*.
[40] A. Zadeh, T. Baltrušaitis, and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," 2017, *arXiv:1611.08657*.
[41] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105.
[42] K. Khan, M. Mauro, and R. Leonardi, "Multi-class semantic segmentation of faces," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 827–831.
[43] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.
[44] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. UM-CS-2010-009, 2010.
[45] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Nov. 2011, pp. 2144–2151.
[46] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.
[47] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "Img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7617–7627.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[49] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5203–5212.
[50] S. Brahnam, L. Nanni, and R. Sexton, "Introduction to neonatal facial pain detection using common and advanced face classification techniques," in *Advanced Computational Intelligence Paradigms in Healthcare–1*. Cham, Switzerland: Springer, 2007, pp. 225–253.
[51] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack, "SVM classification of neonatal facial images of pain," in *Proc. Int. Workshop Fuzzy Log. Appl.*, vol. 3849. Berlin, Germany: Springer, 2005, pp. 121–128, doi: 10.1007/11676935_15.
[52] Z. Tang, X. Liu, H. Chen, J. Hupy, and B. Yang, "Deep learning based wildfire event object detection from 4K aerial images acquired by UAS," *AI*, vol. 1, no. 2, pp. 166–179, Apr. 2020.
[53] X. Zhang, D. Zhang, A. Leye, A. Scott, L. Visser, Z. Ge, and P. Bonnington, "Autonomous incident detection on spectrometers using deep convolutional models," *Sensors*, vol. 22, no. 1, p. 160, Dec. 2022.
[54] Y. Souley Dosso, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D scene analysis in the NICU," *Comput. Biol. Med.*, vol. 138, Nov. 2021, Art. no. 104873.

YASMINA SOULEY DOSSO (Student Member, IEEE) received the B.Sc. degree in biology from Concordia University, Montreal, QC, Canada, in 2014, and the B.Math. degree in statistics and the Ph.D. degree in biomedical engineering from Carleton University, Ottawa, ON, Canada, in 2016 and 2022, respectively (after fast-tracking from master's degree).

Her thesis focuses on machine vision technologies for non-contact video-based neonatal patient monitoring. Her research interests include deep learning and machine vision applications in healthcare, image processing techniques, multimodal data fusion, depth data analysis, health informatics, and patient monitoring system design. Since 2016, she has been an Active Member (the Co-Chair, since 2019) of the IEEE Engineering in Medicine and Biology Society at Carleton University, which has received multiple IEEE awards in recognition of the club's exceptional work.

DANIEL KYROLLOS (Student Member, IEEE) received the B.Eng. degree in biomedical and electrical engineering from Carleton University, where he is currently pursuing the Master of Applied Science degree in electrical and computer engineering with a specialization in data science. His research interests include data science, machine learning, machine vision, natural language processing, and signal processing. Currently, his thesis project aims to investigate novel patient monitoring technologies in the neonatal intensive care unit. The goal is to examine the use of pressure-sensitive mats, video data and depth data to provide continuous, unobtrusive, and non-contact monitoring of critically ill babies.

KIMBERLEY J. GREENWOOD received the Diploma degree in electrical engineering technology from the Ryerson Polytechnical Institute, Toronto, ON, Canada, in 1984, the B.A.Sc. degree in technology management from Bemidji State University, Bemidji, Minnesota, USA, in 2006, and the M.A.Sc. degree in biomedical engineering from Carleton University, Ottawa, ON, Canada.

He is currently a Licensed Professional Engineer in the Province of Ontario and a certified Clinical Engineer. He was recognized as a fellow of the Engineering Institute of Canada, in 2018, and as a fellow of the Canadian Medical and Biological Engineering Society, in 2020. He is also an Adjunct Professor in mechanical engineering with the University of Ottawa and the Executive Director of the Eastern Ontario Clinical Engineering Service and the Chief Clinical Equipment Officer with Children's Hospital of Eastern Ontario, Ottawa.

JOANN HARROLD received the B.Sc. and M.D. degrees from McMaster University, Hamilton, ON, Canada, in 1994 and 1997, respectively.

She holds the Royal College certification in Pediatrics (training with the University of Toronto, Toronto, ON, Canada) and in Neonatal-Perinatal Medicine (training at McMaster University, Hamilton, ON, Canada). She is currently an Associate Professor with the Division Chief of Neonatology, Faculty of Medicine, University of Ottawa, Children's Hospital of Eastern Ontario and the Division Head of Newborn Care at The Ottawa Hospital, Ottawa, ON, Canada.

JAMES R. GREEN (Senior Member, IEEE) received the B.A.Sc. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1998, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from Queen's University, Kingston, ON, Canada, in 2000 and 2005, respectively.

He is currently a Professor with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON. His research interests include machine learning challenges within biomedical informatics, patient monitoring, computational acceleration of scientific computing, and the design of novel assistive devices.