IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 April 2023, accepted 27 April 2023, date of publication 1 May 2023, date of current version 8 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272214

RESEARCH ARTICLE

Mask Detection and Classification in Thermal Face Images

NATALIA KOWALCZYK, (Member, IEEE), MILENA SOBOTKA, AND JACEK RUMIŃSKI, (Senior Member, IEEE)

Department of Biomedical Engineering, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 80-233 Gdańsk, Poland

Corresponding author: Natalia Kowalczyk (natalia.kowalczyk@pg.edu.pl)

This work was supported in part by the IDUB Combating Coronavirus Program "Thermographic system for automatic monitoring of people with increased body temperature" under Grant 14/2020/IDUB/I.3/CC; and in part by the Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdańsk University of Technology.

ABSTRACT Face masks are recommended to reduce the transmission of many viruses, especially SARS-CoV-2. Therefore, the automatic detection of whether there is a mask on the face, what type of mask is worn, and how it is worn is an important research topic. In this work, the use of thermal imaging was considered to analyze the possibility of detecting (localizing) a mask on the face, as well as to check whether it is possible to classify the type of mask on the face. The previously proposed dataset of thermal images was extended and annotated with the description of a type of mask and a location of a mask within a face. Different deep learning models were adapted. The best model for face mask detection turned out to be the Yolov5 model in the "nano" version, reaching mAP higher than 97% and precision of about 95%. High accuracy was also obtained for mask type classification. The best results were obtained for the convolutional neural network model built on an autoencoder initially trained in the thermal image reconstruction problem. The pretrained encoder was used to train a classifier which achieved an accuracy of 91%.

INDEX TERMS Deep neural networks, epidemic prevention, health infrastructure, mask area detection, mask type classification, thermal imaging.

I. INTRODUCTION

Due to the emergence of the coronavirus pandemic in the world, wearing face masks is no longer a novelty, not only in the case of this one disease. Many solutions are based on assessing whether a face mask has been worn - which is essential when epidemiological restrictions apply, for example, when monitoring entrances to buildings and hospitals. Wearing masks allows for the reduction of the spread of diseases, including COVID, influenza, etc.

Machine learning algorithms, in particular deep learning, can be used to solve the classification problem - of determining whether a face mask is worn or not. In [1], the authors proposed a Deep Masknet model that can be used to detect a mask on a face (actually perform the binary classification: "mask", "no mask"). The proposed model for the classification task was verified using the Facemask [2] dataset, Facemask Detection Dataset (20,000 Images) [3], and for the set FaceMask Dataset [4] achieving accuracy, precision, recall, and F1-score at least 97.5% for each metric. The authors have also developed their own dataset - MDMFR, containing over 6000 RGB images. The classification results obtained for the new dataset were characterized by 100% accuracy.

Authors of [5] proposed a classification model suitable for working with real-time images. The model architecture was based on five convolutional layers, five pooling layers, and one fully-connected layer for classification. It was trained using the Face Mask Detection Dataset [6]. The obtained results indicate the high accuracy of the proposed solution (98%).

In another work, [7], a deep learning model was proposed based on the AlexNet model [8]. Two datasets were used for training: the Real-World Masked Face Dataset (RMFD) [9], and Celeb Faces Attributes (CelebA) [10]. The study used the pixel-oriented algorithm with a Deep

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

C2D-CNN (color 2-dimensional principal component analysis (2DPCA)-convolutional neural network) model to detect a face.

A model based on ResNet50V2 was used to classify faces with or without a mask in [11]. Evaluation of the model on the MAFA [12] set showed accuracy at 90.49%, higher than the other tested base models. The proposed model was optimal regarding inference time, error rate, detection speed, and memory usage among the compared models.

The article [13] proposes detecting three conditions for wearing a mask: correctly, incorrectly, and not wearing it. Using the Labeled Faces in the Wild [14] dataset and applying different mask types on faces, the authors achieved a 92% classification accuracy for the Resnet50 model.

The previously mentioned challenges for masked face images are solved for visible light images. Many related datasets have been proposed. However, only limited datasets are available in other domains, like infrared imaging. Thermal imaging is potentially desirable since it can provide images even in low-light conditions. Additionally, thermal images are usually represented by less recognizable biometric features and therefore could be more acceptable regarding privacy aspects. Some datasets with thermal face images are also available. One of the most popular databases of facial thermal images is the dataset proposed in [15]. It contains high-resolution images with a wide range of head positions and a high variation of facial expressions. Images have been recorded from 90 people and manually annotated.

The face mask classification problem has also been investigated for thermal images. In [16] analyzed face detection of people wearing masks using images obtained from different types of thermal cameras (with different resolutions and quality of images). Several deep learning models were adapted and verified, showing the ability to detect faces with masks using the Yolov3 model, achieving an mAP of 99.3%, while the precision was at least 66.1%.

A similar classification problem was described in [17]. The model based on MobileNetV2 was used for feature extraction from a thermal image and for detecting if a person is wearing a mask. The private dataset was used with images of size 80 × 60 pixels. The obtained accuracy of determining whether a person is wearing a mask was 98%.

In the article [18], face detection was performed based on features extracted by Max-pooling and fast PCA, and SVM was used to classify these features. The authors relied on a small dataset (containing only 800 images), and the average face mask recognition proposed by the method can be up to over 99.6%. Facial recognition in thermal images was taken up in the article [19]. Face recognition is performed using temperature information. The feature vector underlying the classification consists of the most representative thermal points on the face, and random forests were used as the classification method. The study also considered images with noise and various types of occlusions.

In [20], the authors proposed a network to detect and capture the temperature of a specific point inside a predicted region. They additionally used RGB data for the ResNet50-based RetinaNet model [21] to classify data into 3-classes evaluating how the facial mask is worn: ''good,'' ''bad,'' and ''none.'' The proposed method achieved an average confidence score of 81.31%. They also described problems with head detection accuracy.

Many other studies were focused on the processing of face images with masks. For example, the analyzed problems addressed face recognition (e.g., [22], [23]) or emotion recognition (e.g., [24], [25]) using face images covered by masks.

However, to our knowledge, no studies were published on face mask detection problems in the thermal domain, i.e., localization of a mask within a face. Single studies focus on detecting the location of the mask on the face for visible light images. The authors of [26] have created a face mask detection dataset (FMD) containing over 52,000 images and annotations for class labels, with and without a mask, mask incorrect, and mask area. They proposed a solution based on the YOLOv4 [27] model to detect the position of the mask on the face, achieving an average precision with a value of 87.05%. In another paper from the same research group [28], the ETL-YOLO v4 model was proposed for the detection of various variants of the position of the mask on the face and the detection of the mask area, which was trained and evaluated using the FMD set [26]. The YOLOv4 model in the ''tiny'' version was improved by adding a dense SPP network, two extra YOLO detection layers, and using the Mish activation function. On the test set, it achieved an average precision of mask location detection of 86.97%, while on the whole set, mAP was 67.64%.

Additionally, only limited works addressed the problem of mask type classification. In [29], in addition to the well-known task of classification - whether a person is wearing a mask or not, authors also proposed a classification of the type of mask. Types of masks have been divided into two categories - qualified masks (N95 masks and disposable medical masks) and unqualified masks (mainly including cloth masks and scarves). The authors showed a method based on transfer learning, using the MobileNet [30] model, which achieved an accuracy of 97.84%.

Using thermal imaging for mask recognition under epidemiological restrictions could provide additional information. Analyzing the average temperature change in the face mask region in a sequence of thermal images can be potentially used to estimate the respiratory pattern and rate. In [31], the authors show the visualization of exhalation flows in thermal images while wearing protective face masks. However, the analyzed area is not searched automatically.

In this study, we focused on two main goals: 1) to detect a face mask within a face region of an image and 2) to classify the protective mask type.

The problem of the automatic detection (i.e., localization) of masks on thermal face images is complex. There are no

public datasets of thermal face images with masks. Additionally, thermal images are usually more smooth than visible light images of faces. Therefore, it is much more challenging to distinguish characteristic features of protective masks about the skin in thermal images (facial mask temperature changes towards the skin's surface temperature and it is much more difficult to distinguish the mask's borders). No earlier studies have presented results in this area, so no models are specialized in detecting the location of masks that could be used in the comparison.

The motivation for using thermal imaging is to use the same modality as for a person's body temperature estimation to see what other important information we can obtain. Many methods have been previously proposed to estimate respiratory rate and patterns from sequences of thermal images recorded for the face (nostrils and mouth). During the pandemic, people wear facial masks, so nostrils and mouth are covered. By detecting facial mask areas in thermal images, we can obtain information about the local temperature change caused by breathing. This can be verified in a separate study. First, the mask detection method should be proposed and validated to verify if it is possible to reliably extract mask area from thermal images of faces (mask area detection not face area detection). This is an important novelty of the paper. Proper detection of mask location (not face) gives many possible future applications. This includes extraction of respiratory-related signals (average temperature within a detected mask region for each frame produces the estimated respiratory signal) and verifying if the mask covers the mouth/nostrils area properly. These two methods are not presented in this paper but are described as our motivation and possible future applications.

This work aims to find and train a model that automatically detects the mask's position on the face. We also check whether it is possible to classify the type of mask worn in thermal images. Different models were analyzed for mask detection using the created database of thermal images of people with masks. Classification of the type of face mask was carried out by validating various models using a subset of images.

The main contributions of this work include: 1) Creation of an extended dataset containing over 9,000 images recorded with different types of thermal cameras with different resolutions, showing people in three types of masks. 2) Demonstrating, probably for the first time, that adapted object detection deep models could efficiently localize virus protective masks within a face thermal image. 3) Demonstrating, probably for the first time, that the deep, autoencoder-based model can be successfully used to classify the type of face mask in thermal images.

The paper is structured as follows: in the following section, we first introduce the dataset used in this paper. In section III, we introduce the detail of the model's testing scenario and characterize the details of the models used both for mask classification and detection tasks. Following this, we provide results and a discussion of the obtained results. In section V, we present the conclusions.

FIGURE 1. Examples of images included in dataset with marked mask regions.

II. DATASETS

A. FACE WITH MASK THERMAL DATASET

As no public face mask databases are available, we decided to create our own dataset - Face with Mask Thermal Dataset (FMT Dataset). Therefore we extended a dataset created in our previous work [16] - a dataset consisting of almost 8,000 thermal images showing people's faces (92% of the images were masked), with different quality and different poses of people (real situations were depicted, e.g., entrancing the buildings). Additional images were collected using a FLIR Boson camera (60 fps). Participants put on face three types of masks (an FFP2 mask, a surgical mask, and a cloth face mask) and performed head movements (side-to-side and up-and-down movements) approximately 80 cm from the camera. Every 20th frame from the recording was selected for the dataset. The experiment was performed with permission of the local Committee for Ethics of Research with Human Participants of 02.03.2021. Each of the participants in the experiment gave informed consent to its performance.

The extended dataset includes 9,394 images with new descriptions that describe the position of the mask in the image. In all of the images, people are wearing a mask of various types: a surgery mask, an FFP2 mask, or a cotton face mask. The number of labeled masks in the dataset is 12,306 - there was more than one person in some images. Figure 1 shows examples from the data set with a mask bounding box.

The collected images were recorded using three different cameras (Table 1). The dataset was divided into the training subset (90%) and the test subset (10%). In each of the separated subsets, there are images taken by each camera, and in each of the subsets, there were images of different people. The images were manually labeled using the same software reported in [16]. The criteria for annotating the face mask were: 1. marking the regions that include the whole mask and 2. a region could be annotated if a minimum of 50% of its area was visible. The annotations of masks were made by six people and were checked twice for accuracy and correctness. A subset was extracted from the dataset, which allows the classification of the type of masks into three classes. This

TABLE 1. Descriptions of cameras.

Camera	Spatial resolution	Dynamic range	Frame rate	Number of images in	
				train set	test set
FLIR Systems A320G	320 x 240	16 bit	60 fps	4040	358
FLIR Systems A655SC	640 x 480	16 bit	50 fps	879	83
FLIR Systems Boson	640 x 512	14 bit	60 fps	3511	523

FIGURE 2. Example of three types of mask: (a) FFP2 mask, (b) surgical mask, and (c) cloth face mask.

subset contains 1841 images depicting ten people. It was divided into a training set of 1285 images (from 7 people) and a test set of 556 images (from 3 people). In Figure 2, examples of images of one person in each of the three types of masks were used to classify the type of mask.

B. SIMULATED DATASET

Due to the lack of available databases of thermal images with mask annotations and the number of available thermal images in our collection, we decided to use transfer learning to train mask detection models. All models were first trained on the WIKI dataset (with cropped faces), derived from the IBMD-WIKI dataset [32], which was prepared for mask detection by randomly applying one of eight types of masks to the images. Among the applied masks were drawing masks and masks extracted from thermal images. The [33] tool was used to put the mask on the face in the correct orientation - images with the masks applied and the coordinates of their location were saved. The images were then converted to grayscale to make them similar to thermal images. Figure 3 shows sample images from the WIKI collection, masked and converted to grayscale. Masks were applied only to images where a face was detected. The obtained set was divided in a ratio of 9:1 into a training set and a test set.

III. METHODS

A. ADAPTATION OF DEEP LEARNING MODELS TO FACE MASK DETECTION TASK

After the extended state-of-the-art analysis, we decided to adapt two models that are the efficient solution for detecting

FIGURE 3. Example of four of the eight types of masks added to images from the WIKI collection and converted to grayscale.

TABLE 2. Models hyperparameters.

Model name	Base model	Number of epochs	Batch size	Optimizer	Initial learning rate
RetinaNet	ResNet-18	100	32	SGD	0.0001
	ResNet-101	100	32		
Yolov5 nano	-	150	32	SGD	0.001

in visible light images. The architectures with a small number of parameters of the considered models were selected because of the limited number of available thermal images with facial masks.

The first adapted model was the nano Yolov5 [34]. The Yolov5 model was created for object detection and can be easily extended to custom data. The "nano" version of the adapted Yolov5 model has 1.9M trainable parameters in total. In comparison, the "small" version has 7.2M parameters. Model training approaches were used with or without transfer learning. It is described later in this section.

The second model chosen in this study was RetinaNet [21]. As the backbone, the ResNet model [35] with 18 layers was selected for calculating the feature maps due to the smallest number of parameters. Additionally, we decided to check another backbone - the ResNet-101 model, which contains a more significant number of layers and will allow comparing the impact of the number of parameters on the metric values obtained during face mask detection. This model is often used for face detection (e.g., [36], [37]) for visible light recorded images as well as in the domain of thermal images - for example, for human detection (e.g., [38], [39]).

All models were trained using the training hyperparameters presented in Table 2.

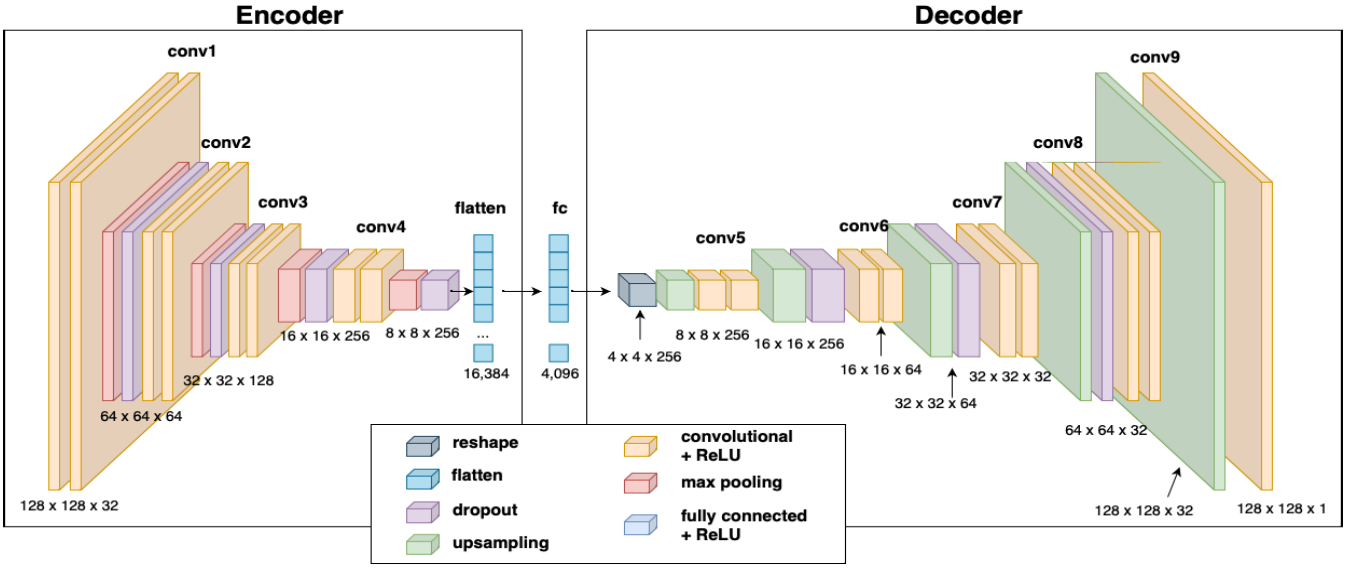


FIGURE 4. Architecture of convolutional autoencoder.

The training was also carried out with or without transfer learning for each model. Two different sets of pretrained initial weights were used: COCO set [40] and WIKI set [32] with masked faces. During the transfer learning scenario, the feature extraction part of the model was frozen. This approach will allow to analyze different scenarios and choose the best model training strategy.

B. DEEP LEARNING MODELS IN FACE MASK CLASSIFICATION TASK

We decided to use a semi-supervised convolutional neural network (CNN) with Convolutional Autoencoder (CAE) as the first phase in the mask classification task. The use of a combination of autoencoders with DNN for a classification problem is commonly used for different tasks, for example, for intelligent fault diagnosis of main reducer [41] or make-up detection [42]. The Autoencoder model was inspired by the [42] model and is used for feature extraction in unsupervised model training using unlabelled data. The weights obtained in the CEA training will be used to initialize the CNN weights in the supervised learning approach. The model architecture used in this study is shown in Figure 4. After each convolutional layer (except the last one), Batch Normalization was applied. The model's training lasted 50 epochs, and Adam was used as the optimizer with a learning rate of 0.00015. The loss function used was binary cross entropy.

Semi-supervised learning scheme for the mask classification task is presented in Figure 5. Using the autoencoder part - the encoder and adding two dense layers on top, with 256 and 128 neurons. Then the softmax function was used for classification into three classes. The trained autoencoder model in the reconstruction task was used in the classification process. The weights for the encoder were used for initialization, and the weights for the classifier part were trained from scratch

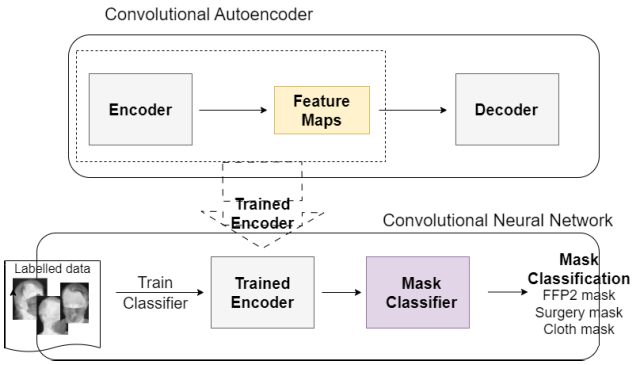


FIGURE 5. Semi-supervised learning scheme for mask classification task.

using labeled data. The face mask classifier was trained by 100 epochs, the batch size was 32, and the optimizer used was Mini Batch Gradient Descent (with a learning rate of 0.001).

Two other models were used to compare the proposed approach with other popular classification models. The first of them was ResNet-50 [35]. At the top, a classification part was added, similar to the CAE-based CNN, consisting of two dense layers (with 256 and 128 neurons, respectively) and a classification layer. The input images were 128 × 128x1. During the model's training, the weights obtained by the model on the ImageNet [43] set will be used, and the classifier will be trained from scratch. Other training parameters will be identical to those for the semi-supervised CNN.

Vision Transformer was proposed as a second architecture to compare with the CAE-based CNN model. A model designed to work with small sets of data [44] was used, which uses the Shifted Patch Tokenization (SPT) block. A dropout layer has been added between the SPT block and the Transformer. For the proposed model, the parameters presented

TABLE 3. Vision Transformer model parameters.

Parameter	Value
Number of patches (patch_size)	8
Size of the output tensor after the Linear layer (dim)	512
Number of Transformer blocks (depth)	4
The number of heads in Multi-head Attention layer (heads)	8
FeedForward layer size (mlp_dim)	512
Dropout rate (dropout)	0.1
Dropout rate for Embedding (emb_dropout)	0.1

in 3 Table were used. During the training of the model for 100 epochs, Adam with a learning rate of 0.00003 was utilized as the optimizer, and a batch size was 16. Cross entropy was used as the loss function. In addition, data augmentation consisting of random horizontal flips and crops of a random portion of the image was used to prevent overfitting.

Classifying the type of masks was carried out using a separate subset allowing for the classification of masks on the face. To prepare the images for the classification model training, they were subjected to preprocessing, which consisted of extracting only the face of the person in the image. This will provide the model with a fragment of the image on which it can focus, thus removing unnecessary background elements. To extract faces from the images, the Yolov3 model [45] was used, which was trained to detect faces of people with masks in thermal images described in our previous work [16].

IV. RESULTS

A. MASK DETECTION

For all models, each test scenario was repeated three times, and the results are presented as the mean value and standard deviation of the results obtained from single attempts.

Table 4 shows the results obtained for four different training approaches of the Yolov5 model in the ''nano'' version. As can be seen, the highest value of the mAP_{50} metric was obtained when initial weight values were transferred from the model pretrained on the COCO set (RGB images). Only slight differences in precision, recall, and mAP_{50} were obtained for the investigated types of initial weights strategies. High values, i.e., higher than 93%, of quality metrics were achieved in all cases. The repeatability of results for each approach is high; however, the highest standard deviation was obtained for the approach with random initialization of weights as assumed.

The metric values obtained for the RetinaNet model are shown in Tables 5 and 6. Comparing the results obtained for two different base models, an increase in the mAP_{50} and recall value for the base ResNet-101 model is visible for all types of training. The precision value for the model with fewer parameters - ResNet-18 - decreased for most test cases. For both approaches, the results obtained are high, and a model trained in this way could be used in an application that allows the detection of a face mask area. As the RetinaNet model with the highest parameter values, the model trained on a set of thermal images with weights obtained during training on the COCO set, where the ResNet-101 model was the model base, can be indicated. For this model, the standard deviation in training repetitions is lower, which gives a better representation of the results on a small set, despite the more significant number of parameters.

Figure 6 presents the values of losses obtained for the training and test sets during the training of the best versions of Yolov5 and RetinaNet models. Please notice that different loss functions were used in the models. The loss function depicted in the graphs is the bounding box regression loss, showing the difference between the predicted boundary box and the ground truth. For the ResNet-101 based model, the loss function was Smooth L1 loss, while for the Yolov5 model was Complete Intersection over Union function. Analyzing the presented graphs, it can be seen that for both models, the loss values rapidly decreased during the first ten epochs. For

TABLE 4. Results obtained for the Yolov5 model in the nano version on the test set.

Type of training	Precision	Recall	mAP_{50} 50
Training on a thermal images dataset with randomly initialized weights	0.936 ±0.033	0.948 ±0.020	0.964 ±0.025
Training on a thermal images dataset with weights obtained on the COCO set	0.964 ±0.025	0.935 ±0.006	0.970 ±0.013
Training on a thermal images dataset with weights obtained on the WIKI set with masked faces	0.935 ±0.008	0.954 ±0.007	0.966 ±0.009
Training on a thermal images dataset with weights obtained on the WIKI set with masked faces and frozen backbone	0.939 ±0.004	0.932 ±0.008	0.954 ±0.005

TABLE 5. Results obtained for the RetinaNet model with ResNet-18 as a backbone on the test set.

Type of training	Precision	Recall	mAP_{50}
Training on a thermal images dataset with randomly initialized weights	0.962 ±0.023	0.926 ±0.027	0.946 ±0.019
Training on a thermal images dataset with weights obtained on the COCO set	0.964 ±0.008	0.931 ±0.013	0.944 ±0.007
Training on a thermal images dataset with weights obtained on the WIKI set with masked faces	0.967 ±0.008	0.915 ±0.010	0.941 ±0.011
Training on a thermal images dataset with weights obtained on the WIKI set with masked faces and frozen backbone	0.971 ±0.007	0.914 ±0.010	0.944 ±0.012

TABLE 6. Results obtained for the RetinaNet model with ResNet-101 as a backbone on the test set.

Type of training	Precision	Recall	mAP_{50}
Training on a thermal images dataset with randomly initialized weights	0.957 ±0.014	0.936 ±0.012	0.948 ±0.010
Training on a thermal images dataset with weights obtained on the COCO set	0.959 ±0.010	0.941 ±0.006	0.951 ±0.007
Training on a thermal images dataset with weights obtained on the WIKI set with masked faces	0.970 ±0.007	0.924 ±0.008	0.946 ±0.008
Training on a thermal images dataset with weights obtained on the WIKI set with masked faces and frozen backbone	0.965 ±0.008	0.930 ±0.008	0.948 ±0.008

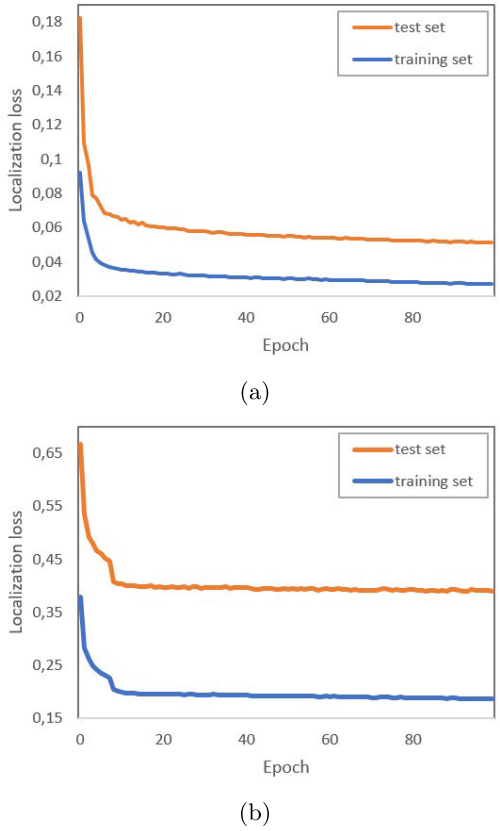


FIGURE 6. Example of loss function change during models training on a thermal images dataset with weights obtained on the COCO set for: (a) Yolov5 model - CIoU loss (b) ResNet-101 based RetinaNet model - Smooth L1 loss.

the test sets losess are slightly larger than in the case of the training set, but they retain the decreasing trend in the training cycle, which proves the correct course of the training.

Examples of mask area detection by the best version of the Yolov5 model and RetinaNet (ResNet-101 based) are shown in Figure 7. For each model, an example of mask position prediction with a high Intersection Over Union (IoU) and a much lower one is shown. The ground truth bounding box was marked in yellow, and the predicted bounding box in blue. The presented detection examples have confidence above 0.9.

B. MASK TYPE CLASSIFICATION

The results obtained for all mask classification models are presented in Table 7. The accuracy value achieved by the CNN based on the CAE model shows that 91% of the images from the test set are correctly classified.

High precision and recall values were obtained for each type of a mask. Analyzing the obtained F1-score values for each of the classes, they illustrate a high balance between precision and recall for each of the classes. In Figure 9 confusion matrixes are presented demonstrating results of CNN based on CAE, Resnet-50 based model and Vision Transformer. The classification model pretrained on the autoencoder correctly classified all examples belonging to "FFP2 mask"

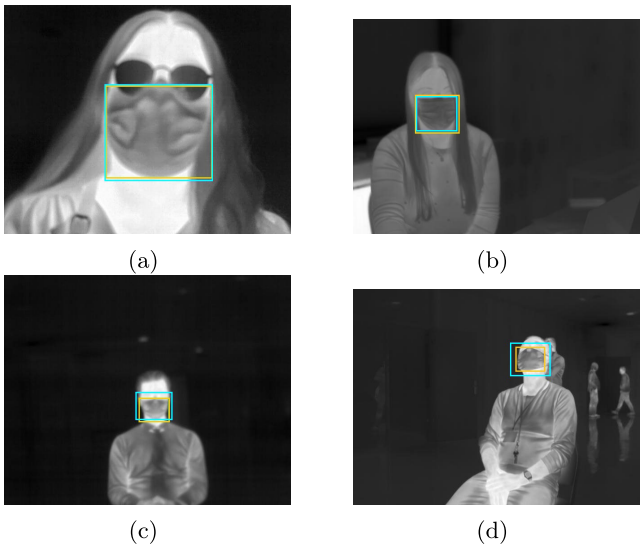


FIGURE 7. Examples of Yolov5 and RetinaNet (ResNet-101 based) results (in blue) vs. ground truth (in yellow). Best matching: (a) Yolov5 - IoU=0.954, (b) RetinaNet - IoU=0.927; Worse matching: (c) Yolov5 - IoU=0.601 and (d) RetinaNet - IoU= 0.525.

TABLE 7. Results obtained for the classification models on the test set.

	Mask type	Precision	Recall	f1-score	Accuracy
CNN based on CAE	Cloth	0.96	0.90	0.93	0.91
	FFP2	0.85	1.00	0.92	
	Surgery	0.96	0.84	0.90	
ResNet-50 based model	Cloth	0.74	0.83	0.78	0.81
	FFP2	0.79	0.92	0.85	
	Surgery	0.98	0.70	0.81	
Vision Transformer model	Cloth	0.93	0.63	0.75	0.85
	FFP2	0.93	0.95	0.94	
	Surgery	0.74	0.96	0.84	

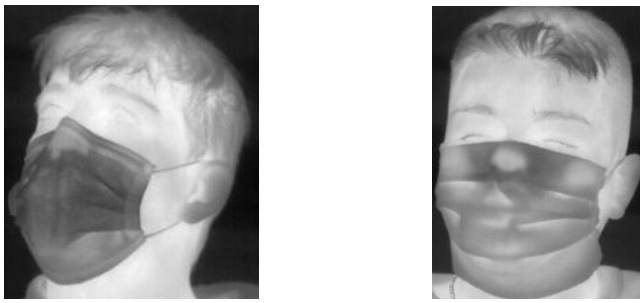
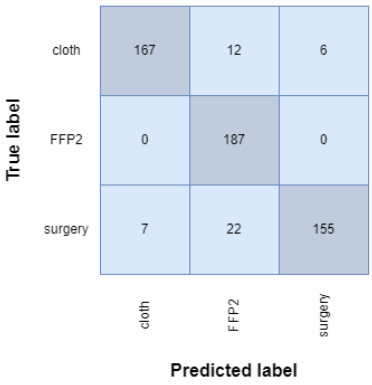


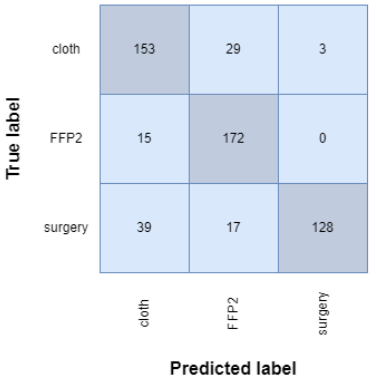
FIGURE 8. Missclassification made by CNN based on autoencoder model: (a) predicted label: "FFP2", true label: "surgery" and (b) predicted label: "surgery", true label: "cloth".

class in the test set. Several incorrect classification results were observed for other two types of facial masks. The most incorrect classification is the assignment of surgery or cloth masks to the FFP2 class. Figure 8 depicts an example of misclassifications made by the CNN based on CAE model.

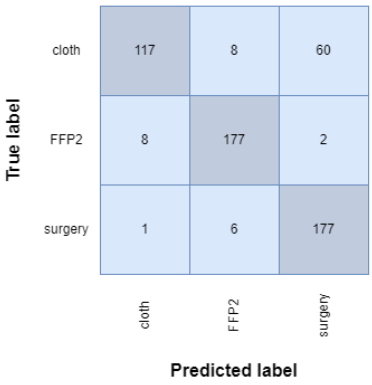
The accuracy of mask classification for the ResNet-50 based model was 81% which is good but much lower than for the CAE-based model. Analyzing the confusion matrix, a more significant number of mistakes is observed for this



(a)



(b)



(c)

FIGURE 9. Confusion matrices for the classification models on the test set: (a) CNN based on CAE, (b) ResNet-50 based and (c) Vision Transformer.

model. The most challenging task for the ResNet-50 based model was correctly classifying surgery and cloth based masks. Again the classification results for "FFP2 masks" are the best. The highest precision is obtained for surgery masks, and there were only three wrong assignments of cloth masks for this class.

The Vision Transformer (VT) model results are also worse than for CNN based on CAE. Comparing results for all models, the value of the F1-score for FFP2 masks is the highest for the VT model. Collating the measures obtained for individual types of masks, this model could be better at correctly classifying cloth masks, achieving a recall of only 63% due to incorrectly assigning them to the surgery class. However, the overall results are the best for CNN based on CAE, showing the high and balanced F1-score values for all types of masks.

The proposed best solutions (weights, code for test, and thermal images examples) for mask detection and classification are available at https://github.com/natkowalczyk/thermal-mask-classification-and-detection.

V. DISCUSSION

Adapting deep neural network models for object detection allows the location detection of facial masks in thermal images. Three models were trained, each in four test scenarios. This made it possible to verify if the results were accidental and compare the models. Additionally, it was possible to check whether transfer learning would allow for better results than training the model from scratch or fine-tuning it. Facial masks appear differently in thermal images than in visible light images. For example, the appearance depends on the breathing phase that modifies the temperature distribution at the observed mask surface. The appearance of facial masks in RGB images does not depend on physiological phenomena. Additionally, it is much easier to obtain or synthesize RGB images with facial masks (e.g., [26]). Therefore, theoretically, transfer learning could be used to reuse the model's weights obtained during training with visible light images as freeze or initial weights in training a model with thermal images. The results showed that using weights pretrained on the COCO set (no masks) as initial weights led to the best localization precision after the proper training on thermal images. However, the maximum difference between analyzed strategies was mAP_{50}=2.9%, precision=3.6%, and recall=4%. The Yolov5 model ("nano" version) gave the best results mAP_{50}=97%, precision=96.4% and recall=93.5%. The "nano" version of the Yolov5 model was experimentally chosen because it produced the best results and due to the smallest number of parameters, which allowed the reduction of the overfitting problem. Other methods like early stopping and image augmentation (e.g., image rotation, flipping) were used to reduce overfitting. Different types of thermal images were also used to properly generalize the data (different resolutions and different image quality). Instead of cross-validation, many experiments with a random selection of batches were provided.

It is difficult to compare the achieved results to other studies because, to our knowledge, the lack of published papers on facial mask localization within thermal images of the face. Related thermal image datasets are mostly private and more difficult to collect. So, only a few papers have focused on face [46] or masked face [16] detection for such

images. In [16], authors used the Yolov3 model to detect faces with masks in thermal images. The private dataset consisted of instances of two classes: ''mask'' and ''no mask''. The images represented different human poses at different distances from the camera. The obtained the mAP_{50} value was 99.3% and the precision was 66.1%. The low precision was probably caused by a wide variety of low-quality thermal images with masked faces recorded from long distances. In this study, the extended dataset was used with additional images (about 15% more) presenting faces closer to the camera. Therefore, the obtained precision highly improved, reaching 95.9% for the best model, while mAP_{50} was only slightly lower (by about 2%).

Detection of face masks was also proposed in [26] and [28]. However, the authors only focused on visible light images using the FMD database. They investigated the detection and classification problem of face mask images into four classes. The detection of the mask area (one class) achieves an average precision of about 87% for both models, while for all classes mAP was 67.64% for ETL-YOLO v4 and 71.69% for Yolov4 based solution. For our scenarios and models, achieved mAP_{50} is over 94% while retaining high precision and recall values simultaneously.

This study also addressed the problem of facial mask type classification. The proposed CNN model based on a convolutional autoencoder (CAE) architecture achieved the best results in classifying mask types.

To our knowledge, no previous studies have classified the type of mask on the face in thermal images. Additionally, only a limited number of studies have been performed on mask classification in visible light images. In [29], the mask was classified into two classes (qualified and unqualified), and an accuracy of 97.84% was achieved. For our best model - CCN based on CAE, the accuracy was 91%, but the masks were classified into three more specific classes. In addition, it is worth noting that in thermal images, the features are usually smoother and lower quality than in RGB images; therefore, the result achieved by the proposed model is relatively high.

This study is probably the first on face mask classification in thermal images. It could be potentially used in various types of monitoring applications when it is necessary to check the wearing of the correct type of mask.

The interesting observation is the higher recall for FFP2 masks. The test set was well-balanced, so the difference in classification efficiency among other mask types could be caused by the geometry of the FFP2 mask from other types of masks. They resemble a duck's bill, introducing more high-frequency features (edges), which may affect the feature extraction. For surgery or cloth type masks, errors in the erroneous classification of examples within these two classes may be due to the similar shape of these masks. It is also worth noting that in the training and test sets, people's faces are registered at different angles, so in some cases, it may be difficult to distinguish the type of mask.

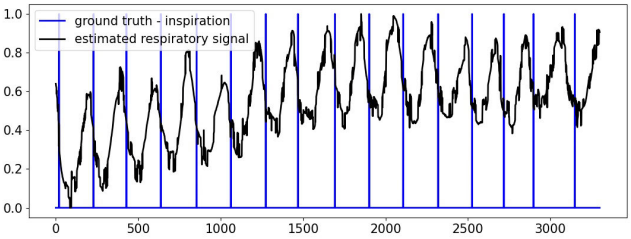


FIGURE 10. Example of the respiratory signal obtained for a sitting person in an FFP2 mask based on detected mask area (average value in a mask area for each frame). A subject indicated the start of the inspiration phase (ground truth).

Thermal imaging is effectively used in the estimation of respiratory rate. In [47], the possibility of using a portable thermal camera to estimate breathing parameters based on a video sequence was presented. It has been shown that the rate and periodicity of respirations can be reliably assessed. Similarly, in [48], Super Resolution Deep Neural Network was proposed, allowing for improving the accuracy of respiratory rate estimation from low-resolution thermal sequences. The topic of determining RR using thermal imaging was also taken up in many other papers, for example, in [49], in the context of monitoring this parameter in newborns. The above works show that using thermal imaging (even from very low-resolution cameras) allows for estimating the respiratory rate. In times of pandemic, when wearing face masks is mandatory, detecting the mask area on the face can probably allow estimating the local temperature change in the area of a mask. It could potentially improve the accuracy of the respiratory rate estimation compared to using the entire face area with only local changes near the nostrils and mouth. Figure 10 shows an example signal for a person wearing an FFP2 (N95) mask, extracted on the basis of the detected area of the mask on the face. Based on the signal shown, the regularity of the breath can be observed and its frequency can be calculated. This signal was estimated for one of the co-author's thermal sequences and is used here only to illustrate possible further studies and applications.

Detection of the position of the mask on the face about the facial feature points may allow checking whether the mask on the face is correctly put on and covers the mouth and nose. This issue is significant about the epidemiological approach presented in [50] - the spread of droplet-borne diseases (when speaking, breathing, coughing, etc.) can be reduced by wearing face masks. Improper wearing of masks (not covering the nose or mouth) does not fully bring the expected results, and the effectiveness of preventing the spread of the disease decreases. In Fig. 11 an example of possible characteristic points detection (68 facial feature points) in reference to the face in a mask is presented. As can be seen, some of these points for the person wearing the mask are covered by it. Identification of these points in reference to the location of the detected mask will allow determining whether the mask is put on correctly (i.e. covers the mouth and nose) or whether it is put on at all. This would be possible using the mask detection

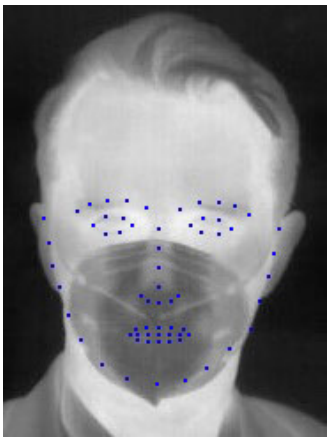


FIGURE 11. Possible facial feature points in reference to the face in a mask. Detected mask's coordinates and detected characteristic facial points can allow determining whether the mask is put on correctly.

method proposed in this paper and will be analyzed in future works.

The issue of classifying the type of mask is advantageous due to the potential of significantly reducing the transmission of SARS-CoV-2 depending on the type of mask worn on the face, shown in [51]. The basic fact is that a properly worn face mask (covering the mouth and nose) can limit the spread of the disease. In addition, a significantly lower virus spread was demonstrated when wearing N95 masks compared to masks used for medical procedures (surgical masks) and cloth masks.

VI. CONCLUSION

It is probably the first study showing that the detection (localization) of face masks in thermal imaging is possible using deep object detection models. Training the models on a prepared and sufficiently large set of thermal images allows for achieving high metric values making this approach potentially interesting for practical applications. For example, the models can be used in further studies to detect if a mask is worn correctly to cover a nose and mouth. Additionally, detecting the mask location on a face can be used to determine the frequency of breathing. It can be achieved by observing a mean temperature change in different phases caused by the breathing process. These problems will be addressed in future studies.

It is also probably the first study that addressed the classification of facial mask types in thermal images. It was shown that the classification of the type of mask worn on the face is possible with relatively high accuracy. For the classification of three types of masks - FFP2, surgery, and cloth, a dedicated CNN model was created based on a convolutional autoencoder. A face mask type classification is useful when requiring a specific type of mask, for example, in some countries, places, etc.

Both aspects of this study, i.e., facial mask localization and mask type classification, can be used together in future applications (e.g., as a part of healthcare infrastructure in hospitals) related to epidemiological screening. It could be important during the epidemic state, pandemic state, or in other related situations (clinics, high environmental pollution, etc.). The use of adequately worn masks and proper mask types can be significant factors in reducing the spread of viruses. This study shows that it is potentially possible to achieve these practical goals by correctly processing thermal recordings. Using thermal imaging can be potentially more acceptable by citizens as it reveals less high-frequency facial features than visible light images and is usually more difficult to match with other personal data.

REFERENCES

[1] N. Ullah, A. Javed, M. Ali Ghazanfar, A. Alsufyani, and S. Bourouis, "A novel DeepMaskNet model for face mask detection and masked facial recognition," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9905–9914, Nov. 2022.

[2] Smansid. (2020). *Facemask Dataset*. Accessed: Nov. 25, 2022. [Online]. Available: https://www.kaggle.com/sumansid/facemask-dataset

[3] J. P. Singaraju and L. Jain. (2020). *Facemask Detection Dataset 20,000 Images*. Accessed: Nov. 25, 2022. [Online]. Available: https://www.kaggle.com/luka77/facemask-detection-dataset-20000-images

[4] S. Shah. (2020). *Facemask*. Accessed: Nov. 25, 2022. [Online]. Available: https://www.kaggle.com/sushantshah/facemask22

[5] H. Goyal, K. Sidana, C. Singh, A. Jain, and S. Jindal, "A real time face mask detection system using convolutional neural network," *Multimedia Tools Appl.*, vol. 81, no. 11, pp. 14999–15015, May 2022.

[6] O. Gurav. (2020). *Face Mask Detection Dataset*. Accessed: Nov. 25, 2022. [Online]. Available: https://www.kaggle.com/datasets/omkargurav/face-mask-dataset

[7] S. Gupta, S. V. N. Sreenivasu, K. Chouhan, A. Shrivastava, B. Sahu, and R. M. Potdar, "Novel face mask detection technique using machine learning to control COVID'19 pandemic," *Mater. Today, Proc.*, vol. 80, pp. 3714–3718, Jan. 2023.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[9] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," 2020, *arXiv:2003.09093*.

[10] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3676–3684.

[11] M. A. S. Ai, A. Shanmugam, S. Muthusamy, C. Viswanathan, H. Panchal, M. Krishnamoorthy, D. S. A. Elminaam, and R. Orban, "Real-time face-mask detection for preventing COVID-19 spread using transfer learning based deep neural network," *Electronics*, vol. 11, no. 14, p. 2250, Jul. 2022.

[12] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2682–2690.

[13] D. Kayali, K. Dimililer, and B. Sekeroglu, "Face mask detection and classification for COVID-19 using deep learning," in *Proc. Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Aug. 2021, pp. 1–6.

[14] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Workshop Faces Real-Life' Images, Detection, Alignment, Recognit., Oct. 2008.

[15] M. Kopaczka, R. Kolk, and D. Merhof, "A fully annotated thermal face database and its application for thermal facial expression recognition," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2018, pp. 1–6.

[16] N. Glowacka and J. Ruminski, "Face with mask detection in thermal images using deep neural networks," *Sensors*, vol. 21, no. 19, p. 6387, Sep. 2021.

[17] B. Sandhya, D. Sesidhar, L. Reddy, T. Meghana, and B. Sony, "Detection of face mask in thermal images using deep CNN," in *Smart Intelligent Computing and Applications*. Berlin, Germany: Springer, 2022, pp. 151–158.

[18] T. Jiang, "GMPPS based face mask detection of infrared thermal image for ultra-small training data set," in *Proc. IEEE Int. Conf. Data Sci. Comput. Appl. (ICDSCA)*, Oct. 2021, pp. 56–60.
[19] S. D. Lin, L. Chen, and W. Chen, "Thermal face recognition under different conditions," *BMC Bioinf.*, vol. 22, no. 5, pp. 1–17, Nov. 2021.
[20] I. Farady, C.-Y. Lin, A. Rojanasarit, K. Prompol, and F. Akhyar, "Mask classification and head temperature detection combined with deep learning networks," in *Proc. 2nd Int. Conf. Broadband Commun., Wireless Sensors Powering (BCWSP)*, Sep. 2020, pp. 74–78.
[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[22] H. N. Vu, M. H. Nguyen, and C. Pham, "Masked face recognition with convolutional neural networks and local binary patterns," *Int. J. Speech Technol.*, vol. 52, no. 5, pp. 5497–5512, Mar. 2022.
[23] W. Moungsouy, T. Tawanbunjerd, N. Liamsomboon, and W. Kusakunniran, "Face recognition under mask-wearing based on residual inception networks," *Appl. Comput. Informat.*, Apr. 2022.
[24] R. Magherini, E. Mussi, M. Servi, and Y. Volpe, "Emotion recognition in the times of COVID19: Coping with face masks," *Intell. Syst. Appl.*, vol. 15, Sep. 2022, Art. no. 200094.
[25] R. Khoeun, P. Chophuk, and K. Chinnasarn, "Emotion recognition for partial faces using a feature vector technique," *Sensors*, vol. 22, no. 12, p. 4633, Jun. 2022.
[26] A. Kumar, A. Kalia, K. Verma, A. Sharma, and M. Kaushal, "Scaling up face masks detection with YOLO on a novel dataset," *Optik*, vol. 239, Aug. 2021, Art. no. 166744.
[27] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[28] A. Kumar, A. Kalia, and A. Kalia, "ETL-YOLO v4: A face mask detection algorithm in era of COVID-19 pandemic," *Optik*, vol. 259, Jun. 2022, Art. no. 169051.
[29] X. Su, M. Gao, J. Ren, Y. Li, M. Dong, and X. Liu, "Face mask detection and classification via deep transfer learning," *Multimedia Tools Appl.*, vol. 81, no. 3, pp. 4475–4494, Jan. 2022.
[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[31] E. Koroteeva and A. Shagiyanova, "Infrared-based visualization of exhalation flows while wearing protective face masks," *Phys. Fluids*, vol. 34, no. 1, Jan. 2022, Art. no. 011705.
[32] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 10–15.
[33] P. Bhandary. (2020). *Mask Classifier*. Accessed: Jun. 3, 2022. [Online]. Available: https://github.com/prajnasb/observations
[34] G. Jocher, "YOLOv5 by ultralytics," Version 7.0, GPL-3.0, May 2020. [Online]. Available: https://github.com/ultralytics/yolov5, doi: 10.5281/zenodo.3908559.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] F. Zhang, X. Fan, G. Ai, J. Song, Y. Qin, and J. Wu, "Accurate face detection for high performance," 2019, *arXiv:1905.01585*.
[37] D. Mamieva, A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Improved face detection method via learning small faces on hard images based on a deep learning approach," *Sensors*, vol. 23, no. 1, p. 502, Jan. 2023.
[38] H. Zhou, M. Sun, X. Ren, and X. Wang, "Visible-thermal image object detection via the combination of illumination conditions and temperature information," *Remote Sens.*, vol. 13, no. 18, p. 3656, Sep. 2021.
[39] T. Hinzmann, T. Stegemann, C. Cadena, and R. Siegwart, "Deep learning-based human detection for UAVs with optical and infrared cameras: System and experiments," 2020, *arXiv:2008.04197*.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf.* Zurich, Switzerland: Springer, 2014, pp. 740–755.
[41] Q. Ye and C. Liu, "An unsupervised deep feature learning model based on parallel convolutional autoencoder for intelligent fault diagnosis of main reducer," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–12, Sep. 2021.
[42] T. Alzahrani, B. Al-Bander, and W. Al-Nuaimy, "Deep learning models for automatic makeup detection," *AI*, vol. 2, no. 4, pp. 497–511, Oct. 2021.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[44] S. Hoon Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*.
[45] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[46] G. Silva, R. Monteiro, A. Ferreira, P. Carvalho, and L. Corte-Real, "Face detection in thermal images with YOLOv3," in *Proc. 14th Int. Symp. Vis. Comput.* Lake Tahoe, NV, USA: Springer, Oct. 2019, pp. 89–99.
[47] J. Ruminski, "Analysis of the parameters of respiration patterns extracted from thermal image sequences," *Biocybernetics Biomed. Eng.*, vol. 36, no. 4, pp. 731–741, 2016.
[48] A. Kwasniewska, J. Ruminski, and M. Szankin, "Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks," *Appl. Sci.*, vol. 9, no. 20, p. 4405, Oct. 2019.
[49] L. Catalina, A. Doru, and C. Calin, "The use of thermographic techniques and analysis of thermal images to monitor the respiratory rate of premature new-borns," *Case Stud. Thermal Eng.*, vol. 25, Jun. 2021, Art. no. 100926.
[50] J. T. Brooks and J. C. Butler, "Effectiveness of mask wearing to control community spread of SARS-CoV-2," *Jama*, vol. 325, no. 10, pp. 998–999, 2021.
[51] B. M. Gurbaxani, A. N. Hill, P. Paul, P. V. Prasad, and R. B. Slayton, "Evaluation of different types of face masks to limit the spread of SARS-CoV-2: A modeling study," *Sci. Rep.*, vol. 12, no. 1, p. 8630, May 2022.

NATALIA KOWALCZYK (Member, IEEE) received the B.Eng. and M.Sc. degrees in biomedical engineering from the Gdańsk University of Technology, Gdańsk, Poland, in 2019 and 2020, respectively. Since 2020, she has been a Teaching Assistant with the Biomedical Engineering Department, Gdańsk University of Technology. Her research interests include machine learning and computer vision for biomedical purposes. She works in a European project named personalized medicine screening and monitoring program for pregnant women suffering from preeclampsia and gestational hypertension and cooperates with companies, carrying out commissioned research in the field of image classification in medicine.

MILENA SOBOTKA received the B.Eng. degree in biomedical engineering from the Gdańsk University of Technology, Gdańsk, Poland, in 2022, where she is currently pursuing the M.Sc. degree. Her research interests include image processing, medical imaging, machine learning, and computer vision.

JACEK RUMIŃSKI (Senior Member, IEEE) received the Ph.D. degree in informatics from the Gdańsk University of Technology, Gdańsk, Poland, in 2002, and the Ph.D. degree in biomedical engineering, in 2016. Since 2022, he has been a Full Professor with the Biomedical Engineering Department, Gdańsk University of Technology, where he is currently the Head. He is the author or coauthor of more than 200 publications and several patent applications. He has been a coordinator or main contractor in about 20 projects, for e.g., eGlasses (CHIST-ERA), WODIA (ERA PerMed), and Ella4Life (AAL). He received a number of awards, including for best papers, medals for practical innovations presented during international fairs, and also the Andronicos G. Kantsios Award. His research interests include medical imaging, image processing, machine learning, human–system interaction, and the quality of life technologies. He is a member of the Committee on Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences, the Chair of the Scientific Committee Chairperson of the Polish AI Society, and a member of the Human-Factor Committee of IEEE IES.