FMR-DBv2: an Improved Database for Mask and Respirator Type and FFP Protection Level Recognition Through Deep Learning

FMR-DBv2: an Improved Database for Mask and Respirator Type and FFP Protection Level Recognition Through Deep Learning

Antonio Costantino Marceddu*, Nicola Dilillo*, Luigi Di Sergio*, Pietro Ruiu†, Andrea Lagorio†, Filippo Casu†, Enrico Grosso†, Renato Ferrero*, and Bartolomeo Montrucchio*

*Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
Email: {antonio.marceddu, nicola.dilillo, luigi.disergio, renato.ferrero, bartolomeo.montrucchio}@polito.it

†Department of Engineering
Università degli Studi di Sassari
Sassari, Italy
Email: {pruiu, lagorio, fcasu1, grosso}@uniss.it

***Abstract*—The widespread adoption of masks and respirators has significantly influenced various aspects of society, driving technological advances to improve comfort, efficiency, and sustainability. The COVID-19 pandemic underscored their essential role in the protection of public health, with continued relevance in the industrial, environmental, and hygiene-critical sectors. Recent developments in deep learning offer promising approaches for building automated systems that can detect mask and respirator usage. In this regard, this paper first aims to present an improved version of the Facial Masks and Respirators Database (FMR-DB), which can be used to create such systems. New features include a significant increase in available images, which has been expanded from 2565 to 4200 images, and the addition of You Only Look Once (YOLO), PASCAL Visual Object Classes (PascalVOC), and Common Objects in Context (COCO) labeling for image detection tasks. Furthermore, image classification and object detection tests were conducted using Convolutional Neural Networks (CNNs), Transformers, and YOLO to determine the types of masks and respirators accurately. Finally, to the best of the authors' knowledge, these tools were used for the first time to analyze the protection levels of respirators automatically. The results provide valuable insights for developing efficient and reliable automatic recognition systems.**

***Index Terms*—Computer Vision, Deep Learning, Image Classification, Image Databases, Object Detection, Personal Protective Equipment**

I. Introduction

The use of Personal Protective Equipment (PPE), particularly those related to respiratory protection, such as masks and respirators, has profoundly impacted many aspects of our society. From a socio-economic point of view, the widespread introduction of these devices has changed daily habits and influenced cultural and social elements, while simultaneously stimulating technological innovation in the personal protection sector. In fact, the prolonged use of masks and respirators can cause discomfort, difficulty breathing, and fatigue. They can also hinder communication in environments where non-verbal language is essential. The resulting drive for innovation has led to continuous improvements that make these devices more efficient and comfortable. At the same time, the search for reusable and easily recyclable solutions continues to be crucial to reduce their ecological impact, since most masks and respirators are made primarily of synthetic polymers.

During the COronaVIrus Disease 19 (COVID-19) pandemic, these devices proved essential to limit the spread of the virus, helping to protect public health [1]. Their function has remained important even during non-pandemic times for reducing the risk of respiratory infections, such as the flu [2]. In the industrial and environmental sectors, respirators equipped with advanced filters, such as the Filtering FacePiece (FFP) 2 and 3, which comply with the European standard EN 149, but also N95 or N99, which comply with the US standard NIOSH-42CFR84, or even KN95 or KN99, which comply with the Chinese standard GB2626-2006, are essential for protecting workers from particles, chemicals, and toxic agents [3]. In addition, they prevent product contamination in the food and pharmaceutical sectors, ensuring high hygiene standards [4]. However, their effectiveness depends on proper and consistent usage. In the IT field, one of the most interesting possibilities offered by the most recent deep learning techniques is the creation of automatic systems capable of:

- Detect whether a person is wearing a mask or respirator; this activity is referred to as *mask detection* problem.
- Recognize the type of mask or respirator worn; this activity is referred to as *mask recognition* problem.

These systems can be created using datasets that differentiate between different types of masks and respirators. The current work aims to contribute in this direction. First, it presents an improved version of the Facial Masks and Respirators Database (FMR-DB). Such a dataset was released in 2021 with the intention of addressing the previously mentioned problems. The improved version presents the following additions:

- The number of images was increased from $2\,565$ to $4\,200$.
- You Only Look Once (YOLO) [5], PASCAL Visual

Object Classes (PascalVOC) [6], and Common Objects in Context (COCO) [7] image labeling were added.
- Many low-quality images have been removed and replaced with higher-quality ones.
- The protection level labels, which were intended for future use in the original version of the dataset, have been revised, and new ones have been added.

Second, some interesting tests will be presented regarding its use to recognize both the typology of masks and respirators and the protection level of the latter. These tests were divided into image classification tests and object detection tests. The former compare the results obtained using Convolutional Neural Networks (CNN) and Transformers, while the latter are performed using YOLO. To the best of our knowledge, no previous work has addressed the prediction of the protection level of disposable respirators.

The paper is organized as follows. Section II discusses some of the relevant works carried out in recent years, while Section III describes the improved FMR-DB dataset. This is followed by a discussion of the evaluation protocols, reported in Section IV, and of the baseline experiments in Section V. The experimental results are reported in Section VI, and finally, the conclusions are discussed in Section VII.

II. Related Work

Some mask detection and recognition systems already existed before the coronavirus: they were used for specific cases, such as monitoring polluted areas [8], for facial detection or recognition [9], and in automatic security systems [10].

During the COVID-19 pandemic, demand for these types of systems has exploded. However, this demand has clashed with the scarcity of databases that could allow their implementation. For this reason, the world of research and industry has moved to create new ones. In March 2020, a ready-to-be-trained facial mask classifier, based on a ResNet-101 Artificial Neural Network (ANN) [11], was released on GitHub [12]. In the self-made dataset of $11\,376$ images, there were 690 images of people wearing an artificially applied surgical mask. This dataset has been exploited in some articles in order to train and evaluate several ANN-based mask detection systems, achieving accuracies of $\sim 99\%$ [13], [14]. It was also employed for training multiple machine learning systems [15]: Decision Tree, Support Vector Machine (SVM) [16], MobileNet [17], MobileNetV2 [18], Xception [19], VGG16 [20], and VGG19. Another example of its use came out in March 2021, in order to develop a new system called Thor to detect unmasked personnel in public spaces [21]. The MaskedFace-Net dataset [22], published in November 2020 and derived from Flickr-Faces-HQ3 (FFHQ) [23], is another dataset dedicated to mask recognition. In this case, however, the masks are artificially superimposed by image processing techniques. MaskedFace-Net also considers a special class to handle cases where the mask is not worn correctly.

These datasets have been useful for creating systems to recognize the presence or absence of a mask on the face. However, artificially applying a mask poses severe limitations due to the poor realism of the images obtained, creating a divergence from the real case [24]. For this reason, newer datasets with real masks have started to appear. In May 2020, Humans In The Loop released the Medical Mask Dataset, which contains $6K$ images acquired from the public domain and covering 20 classes of different accessories, as well as a classification of faces with a mask, without a mask, or with an incorrectly worn mask [25]. In the same month, a system capable of recognizing whether or not a person is wearing a mask was implemented [26]. It discretizes the type of mask worn among the following: homemade, N95, and surgical. This was done using an unpublished, self-made dataset. Moreover, no precise accuracy data are reported, as the trained system was tested on new Internet images and an author's friend, except for a few test images reporting a good performance of the system. On 2021, the FMR-DB database was published [27]. It consists of $2\,565$ images categorized into two classes: one where a mask or respirator is worn, and one where it is not. In the first case, there is a classification between surgical and non-medical masks, as well as between full-face respirators, half-face respirators, and disposable respirators with and without valves. In the second case, there is a classification for the absence or presence of face occlusions. In February 2021, the Face-Mask Label Dataset (FMLD) was released [28]. It was used to train a system capable of recognizing whether a mask is worn correctly (compliant) or worn incorrectly or not (non-compliant), obtaining a recognition accuracy of 98.79%.

Although the use of mask detection and recognition systems was particularly relevant during the COVID-19 pandemic, they continue to play an essential role in improving public health, especially in contexts where it is necessary to prevent the spread of infectious diseases, such as hospitals, public transport, and crowded places, or in high-risk environments.

III. Facial Masks and Respirators Database

A. Previous Version (v1)

The original Facial Masks and Respirators Database (FMR-DB) was created with the intent of enabling the creation of automated systems to classify the type of mask or respirator worn. It was composed of $2\,565$ images retrieved from the Internet, each available in both their original form and a cropped version focusing on the face of the person portrayed. Table I shows the available classes of the dataset and the labels used to name the images in order to indicate the information they contain, while Figure 1 shows an example image for each of the classes contained. The images in the dataset are first divided into the Superclasses *Mask or Respirator*, which contains a breakdown by type of mask and respirator, and *No Mask or Respirator*; the latter is further divided into *No Occlusions*, which depicts images of people without facial occlusions, and *Occlusions*, which instead contains images of people with facial occlusions other than masks or respirators.

Other labels not mentioned in Table I for clarity will be discussed here. The first set of labels refers to the presence or absence of other PPEs in addition to masks or respirators:

- NP $\rightarrow$ *No Eye or Head Protection*.

TABLE I
THE LIST OF LABELS FOR THE FMR-DB DATASET ALONG WITH A COMPARISON OF THE NUMBER OF IMAGES BETWEEN THE PREVIOUS (V1) AND IMPROVED (V2) VERSIONS

Label	Meaning	No. of Images	
		v1	v2
MS	**Mask or Respirator** (Superclass)	**1890**	**2900**
WV	→ Disposable Respirators with Valve	315	500
NV	→ Disposable Respirators without Valve	315	500
FF	→ Full-Face Respirators	315	450
HF	→ Half-Face Respirators	315	450
NM	→ Non-Medical Masks	315	500
SR	→ Surgical Masks	315	500
NN	**No Mask or Respirator** (SuperSuperclass)	**675**	**1300**
NO	→ No Occlusions	315	500
WO	→ ***Occlusions*** (Superclass)	***360***	***800***
HM	→ Hands on Mouth	90	200
HT	→ Hats	90	200
NW	→ Neck Warmers and Bandanas	90	200
SN	→ Sunglasses	90	200

Fig. 1. Example images of each of the classes available in the FMR-DB dataset[1]. The classes available in the dataset are shown in Table I.

- EP → *With Eye Protection.*
- HP → *With Head Protection.*
- EH → *With Eye and Head Protection.*

This set of labels is only present for images in which masks or respirators are present. Another set of labels concerns the protection level offered by *Disposable Respirators*:

- P1 → *FFP1 - N80 - KN80 Protection Level.*
- P2 → *FFP2 - N95 - KN95 Protection Level.*
- P3 → *FFP3 - N99 - KN99 Protection Level.*

This type of label is only available for *Disposable Respirators*. Masks, such as *Non-Medical* and *Surgical*, generally have a lower protection level than FFP1, FFP2, and FFP3, so this labeling cannot be applied. As for *Full-Face* and *Half-Face* respirators, they usually have replaceable filters with advanced levels of protection. However, it is often impossible to distinguish the traits of the filters via image because they are frequently hidden inside the mask itself.

B. Improved Version (v2)

Several additions have been made to the improved version of the dataset. First, the number of images has been increased to 4 200, providing a more comprehensive knowledge base of the different types of masks, respirators, and other facial occlusions that do not fall into the previous two categories. Second, labels in YOLO [5], PascalVOC [6], and COCO [7] formats have been added to enable both image classification and object detection tasks. No labeling has been added for superclasses, since they are only used for discretization between the classes available in the database. Finally, some images have been replaced with higher-quality versions, and certain labels have been refined for greater clarity and conciseness. Since the database is composed of images taken from the Internet, therefore without copyright by the authors, it is possible to request it only by accepting the conditions of fair use [29].

IV. EVALUATION PROTOCOLS

Two protocols with distinct objectives have been developed to assess the dataset and validate its applicability. Each protocol utilizes a unique selection of images organized into representative classes relevant to the studied scenario.

A. Protocol 1 - Mask and Respirator Type Recognition

The first experimental protocol aims to recognize the type of mask or respirator worn. Eight distinct classes have been identified, as reported in Table II, each characterized by specific characteristics.

For this protocol, 4 000 images were selected from the entire dataset. Images labeled as *SN*, corresponding to individuals wearing sunglasses, were excluded as they were considered irrelevant for this scenario. The selected images were divided into two subsets: one for training and one for testing. Specifically, 100 images were randomly chosen from each class for testing, while the remaining 3 200 images were used for training and validation. The original labels of the dataset

[1]In order from top-left to bottom-right, the images were taken from the following sources (accessed on March 25, 2025):

a) https://www.3m.com/3M/en_US/p/d/v000057488/
b) https://www.moldex-europe.com/it/product-detail/compactmask/
c) https://www.gvs.com/it/catalogo/maschera-segre-n99
d) https://izh.technoavia.ru/katalog/siz/siz_dyhaniya/respirator/8-328.htm
e) https://sergios.co.nz/diesel-face-mask-updating/
f) https://www.officecrave.com/kimberly-clark-professional-47080.html
g) https://www.galleriavittoria.com/datrino-elena
h) https://www.freepik.com/free-photo/handsome-man-looking-shocked-speechless-holding-hands-mouth-standing-yellow-background_13082728.htm#from_element=detail_alsolike
i) https://www.aliexpress.com/i/1005005912440782.html
j) https://www.etsy.com/it/listing/803134146/daisies-blu-e-viola-floral-multi
k) https://www.na-kd.com/en/products/sharp-square-cateye-sunglasses-black

TABLE II
THE CLASSES SELECTED FOR PROTOCOL 1 AND THEIR RESPECTIVE MEANINGS

Class	Description	No. of Images		
		Train + Val.	Test	Total
MS_WV	Disposable Respirators with Valve	400	100	500
MS_NV	Disposable Respirators without Valve	400	100	500
MS_FF	Full-Face Respirators	350	100	450
MS_HF	Half-Face Respirators	350	100	450
MS_NM	Non-Medical Masks	400	100	500
MS_SR	Surgical Masks	400	100	500
NN_NO	Person	400	100	500
NN_WO	Hands on Mouth + Hats + Neck Warmers and Bandanas	500	100	600

images are provided in the *Description* column of Table II. The total number of images per class is reported in the *No. of Images* column, specifying the distribution of images used for training, validating, and testing the models.

B. Protocol 2 - FFP Protection Level Recognition

The second protocol focuses on identifying the protection level provided by disposable respirators by classifying the related images into four distinct categories, detailed in Table III. Three of these categories correspond to the protection levels *FFP1*, *FFP2*, and *FFP3*, while the fourth class, defined as *NFPP*, comprises 200 images of masks with no protection level. These images were selected randomly and in a balanced way between those labeled as *NM* and *SR*. A total of 1 200 images were selected, with 1 080 allocated for training and validation and 120 for testing. Due to the limited number of available images, the number of test images was reduced to 30 per class. The details and distribution of images across the classes are provided in Table III. While in Protocol 1, the classes are visually quite distinguishable, in the second, the component images of the various classes are pretty similar. Even for a human being, it is not easy to understand the protection level of a respirator. The only way is to read the writing on the respirator when present. Even the colors used for marking are unreliable, as no universal standard currently exists for visually indicating a respirator's protection level: each manufacturer uses a different color scheme.

V. BASELINE EXPERIMENTS

The two protocols were validated by assessing the performance of various AI models on two fundamental computer vision tasks: image classification and object detection. For image classification, several models based on Transformers and CNNs were assessed, while for object detection, different versions and sizes of the YOLO model were compared.

The selected models were fine-tuned using 20 epochs. Then, the k-fold cross-validation technique, with $k = 5$, was applied to obtain the validation dataset. In each iteration, four subsets of the training dataset were used for fine-tuning, while the remaining one-fifth was reserved for validation. The objective

TABLE III
THE CLASSES SELECTED FOR PROTOCOL 2 AND THEIR RESPECTIVE MEANINGS

Class	Description	No. of Images		
		Train + Val.	Test	Total
FFP1	FFP1 - N80 - KN80 Protection Level	97	30	127
FFP2	FFP2 - N95 - KN95 Protection Level	631	30	661
FFP3	FFP3 - N99 - KN99 Protection Level	182	30	212
NFFP	Non-Medical + Surgical Masks	170	30	200

was to establish baseline results under the given conditions, based on key evaluation metrics such as Accuracy, Precision, Recall, and F1-Score. For the object detection task, additional metrics were considered, namely mAP50 and mAP50-95, where mAP stands for Mean Average Precision.

A. Image Classification

Image classification generally refers to the task of assigning a class name to an entire image. For this task, six distinct families of image classification models were considered: four based on Transformers [30] and two on CNNs [31]. The most suitable models were selected from each family based on their size, defined by the number of parameters in the model, and the intended application. Models of varying sizes from these families underwent preliminary fine-tuning to identify those with the highest accuracy within each group. After determining the best-performing model from each family, additional fine-tuning was performed by increasing the number of epochs to enhance classification performance further.

The Transformer-based models considered for the two protocols include Vision Transformers (ViTs), Shifted window Transformers (Swin), Self-Distillation with No Labels (DINOv2) and Contrastive Language-Image Pre-Training (CLIP).

ViTs [30], [32] adapt the Transformer architecture, originally designed for Natural Language Processing (NLP), to image processing. Instead of analyzing entire images simultaneously, ViTs divide an image into smaller patches, which are then converted into linear embeddings and processed as a sequence. Similar to how words function as tokens in NLP, these image patches serve as the fundamental units for visual data processing. In the experiments, three ViT models with different dimensions and characteristics are used: google/vit-large-patch16-224 (vit-large-patch16), google/vit-large-patch32-384 (vit-large-patch32), and google/vit-base-patch16-224 (vit-base-patch16).

The ***Swin*** model was designed to address challenges in adapting Transformers to vision tasks [33]. It computes image representations using a shifting window mechanism, enabling efficient capture of both local and global visual features. Swin Transformer V2 [34] further enhances the original model, improving its effectiveness for various computer vision applications. Three distinct Swin models, characterized by unique sizes and features, were used in the experiments: microsoft/swinv2-large-patch4-window12to16-192to256-22kto1k-ft (swinv2-large), microsoft/swinv2-base-

TABLE IV
EVALUATION RESULTS OF DIFFERENT MODELS FINE-TUNED ON PROTOCOL 1 WITH 3 EPOCHS

Model	Type	Accuracy
swinv2-large	**Transformer**	**0,9363**
swinv2-base	Transformer	0,9050
swinv2-tiny	Transformer	0,8888
clip-large-patch14	**Transformer**	**0,9275**
clip-large-patch14-336	Transformer	0,9225
clip-base-patch16	Transformer	0,9150
clip-base-patch32	Transformer	0,8813
dinov2-large	**Transformer**	**0,9225**
dinov2-base	Transformer	0,9175
vit-large-patch16	**Transformer**	**0,9150**
vit-large-patch32	Transformer	0,9112
vit-base-patch16	Transformer	0,9050
convnextv2-large	**CNN**	**0,9025**
convnextv2-tiny	CNN	0,8787
efficientnet-b5	**CNN**	**0,8275**

patch4-window8-256 (swinv2-base), and microsoft/swinv2-tiny-patch4-window8-256 (swinv2-tiny).

DINOv2 [35] builds on the idea that pretraining on large datasets, a common approach in NLP, can also benefit computer vision. It is designed to generate general-purpose visual features that can be used across different image types and tasks without requiring fine-tuning. The model employs a ViT with one billion parameters and distills it into a series of smaller, more efficient models. Within the experiments, two DINOv2 versions featuring diverse dimensions and traits are applied: facebook/dinov2-large-imagenet1k-1-layer (dinov2-large), and facebook/dinov2-base-imagenet1k-1-layer (dinov2-base).

Finally, ***CLIP*** [36] was developed to evaluate zero-shot learning, i.e., how well models can classify images without additional training. The original experiment utilized a ViT image encoder with 4 CLIP models of distinct characteristics: openai/clip-vit-large-patch14 (clip-large-patch14), openai/clip-vit-large-patch14-336 (clip-large-patch14-336), openai/clip-vit-base-patch16 (clip-base-patch16), and openai/clip-vit-base-patch32 (clip-base-patch32).

For the CNN-based models, EfficientNet and ConvNeXt V2 were selected for comparison.

EfficientNet [37] is a convolutional model that introduces a novel scaling method. This approach uses a single compound coefficient to adjust the model's depth, width, and resolution uniformly, ensuring a balanced and efficient scaling process. This simple yet effective strategy significantly enhances both performance and efficiency. The experimental setup employed google/efficientnet-b5 (efficientnet-b5).

ConvNeXt V2 [38] is a fully convolutional masked autoencoder framework that incorporates a novel Global Response Normalization (GRN) layer. The GRN layer improves inter-channel feature competition and can be seamlessly integrated into the ConvNeXt architecture. This enhancement substantially boosts the performance of pure convolutional networks across various recognition benchmarks. In the experiments, two ConvNeXt of varying sizes were utilized: facebook/convnextv2-large-1k-224 (convnextv2-large), and facebook/convnextv2-tiny-1k-224 (convnextv2-tiny).

A preliminary fine-tuning of the presented models, with the default optimizer, was performed to identify a reduced subset with the best performance based on the accuracy metric obtained on the validation dataset. The following hyperparameters were used to train them:

- Batch size: 16.
- Learning rate: $5e^{-5}$.
- Number of epochs (preliminary fine-tuning): 3.
- Number of epochs (final fine-tuning): 20.

The results of this preliminary phase of model selection are presented in Table IV. For each model family, the model that obtained the highest accuracy was selected for the final evaluation. For such a phase, the models were fine-tuned using the same parameters as in the previous training, with the only change being the number of epochs, which was increased to 20. The metrics used for this analysis include accuracy, precision, recall, and F1-score.

B. Object Detection

Object detection involves identifying the location of an object in an image. To achieve this, the ***YOLO*** [5] model family was used. YOLO is a widely adopted object detection model in various fields whose popularity is due to its fast and efficient performance. In particular, YOLO is optimized for fast real-time image processing by leveraging a single ANN to predict bounding boxes and classify images simultaneously. This approach simplifies the training pipeline and reduces the overall complexity of the system. Furthermore, its holistic analysis of entire images enables strong generalization across diverse datasets and object categories, enhancing its adaptability to various applications and environments. Multiple versions of the YOLO architecture have been developed over time, each offering unique advantages. In this article, the latest version of YOLO available at the time of writing, i.e., the 12th [39], and the previous one (the 11th [40]), were used in the nano (n), small (s), and medium (m) versions.

The following hyperparameters were used to fine-tune the different YOLO models:

- Batch size: 16.
- Learning rate: $4.76e^{-4}$.
- Number of epochs: 20.
- Optimizer: AdamW [41].

Data augmentation techniques were also adopted to increase, as much as possible, the generalization capacity of the network. The best-performing models were mainly chosen based on the mAP50-95 metric, which is the average of the mean Precision calculated at various Intersection over Union (IoU) thresholds, ranging from 0.50 to 0.95 in increments of 5%. This metric offers a comprehensive assessment of the performance of the model at multiple levels of detection difficulty.

TABLE V
EVALUATION RESULTS OF BEST CLASSIFICATION MODELS FINE-TUNED ON PROTOCOL 1 WITH 20 EPOCHS

Model	Type	Accuracy	Precision	Recall	F1
swinv2-large	**Transformer**	**0,9400**	**0,9414**	**0,9400**	**0,9401**
vit-large-patch16	Transformer	0,9387	0,9399	0,9387	0,9388
convnextv2-large	CNN	0,9363	0,9361	0,9363	0,9360
dinov2-large	Transformer	0,9313	0,9320	0,9312	0,9313
clip-large-patch14	Transformer	0,9287	0,9300	0,9287	0,9289
efficientnet-b5	CNN	0,9163	0,9204	0,9163	0,9164

VI. EXPERIMENTAL RESULTS

A. Image Classification Task

The results of the fine-tuning for the six models selected for Protocol 1 and Protocol 2 are presented in Tables V and VI, respectively, while Figures 2 and 3 show the confusion matrix obtained by the best model in both experiments. As evidenced by the numerical results, in the case of image classification, the swinv2 model emerges as the most performant. In the case of Protocol 2, a significant decline in Accuracy can be observed in Table VI. This drop is due to both the increased complexity of the task and the smaller number of images used for training. Despite the lower Accuracy, a high level of Precision is observed. This suggests that the models are effective in correctly identifying instances of a particular class. It should be noted that ConvNeXt, a CNN-based model, achieved a performance similar to that of the Transformer models in both protocols. This is particularly interesting, as CNNs are an earlier technology compared to Transformers.

Examining the confusion matrices for Protocol 1 (Figure 2), it is evident that all classes exhibit both high Accuracy and high Precision. For Protocol 2 (Figure 3), the matrix highlights the challenges faced in this scenario, revealing differences in classification performance across classes. Specifically, the FFP2 class shows high Precision but low Accuracy, whereas FFP1 and FFP3 exhibit high Precision and low Accuracy. In contrast, the NFFP class achieves high values for both metrics. This behavior is likely due to the imbalance in the number of images used for training and the inherent difficulty in distinguishing respirators with different protection levels.

TABLE VI
EVALUATION RESULTS OF BEST CLASSIFICATION MODELS FINE-TUNED ON PROTOCOL 2 WITH 20 EPOCHS

Model	Type	Accuracy	Precision	Recall	F1
swinv2-large	**Transformer**	**0,6833**	**0,8235**	**0,6833**	**0,6854**
clip-large-patch14	Transformer	0,6750	0,7970	0,6750	0,6837
convnextv2-large	CNN	0,6667	0,7153	0,6667	0,6567
vit-large-patch16	Transformer	0,6583	0,7682	0,6583	0,6466
dinov2-large	Transformer	0,6417	0,7719	0,6417	0,6369
efficientnet-b5	CNN	0,5417	0,7615	0,5417	0,4753

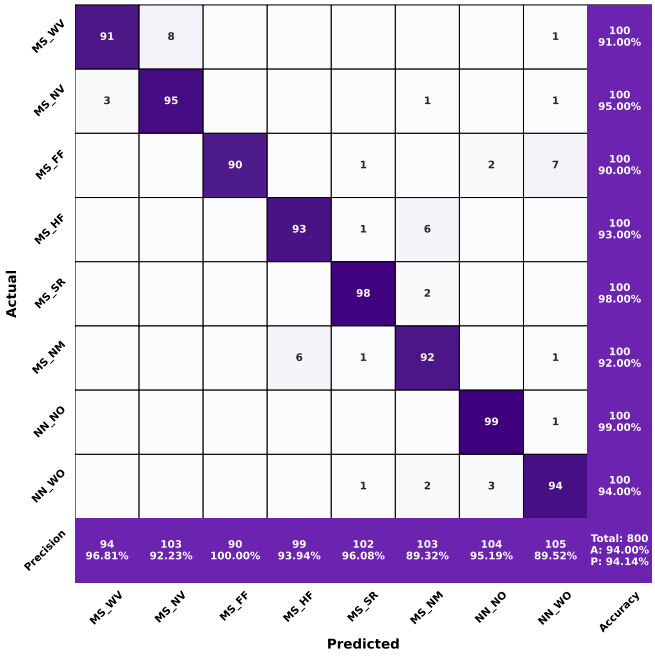

Fig. 2. The confusion matrix obtained by swinv2-large in Protocol 1.

B. Object Detection Task

The results obtained by fine-tuning the different YOLO models for Protocol 1 and Protocol 2 are presented in Tables VII and VIII. For brevity, only the results with the highest mAP50/95 metric are presented, rather than displaying all the obtained results. Generally, the results obtained were relatively consistent between the two tasks, with a 2% difference between the best and worst results for Protocol 1 and a 7% difference for Protocol 2. The networks that performed best for Protocol 1 were YOLO11s and YOLO12s, while for Protocol 2 it was YOLO12n. For consistency with the image classification task, confusion matrices with the same style have also been created for this one, displayed in Figures 4 and 5. However, it

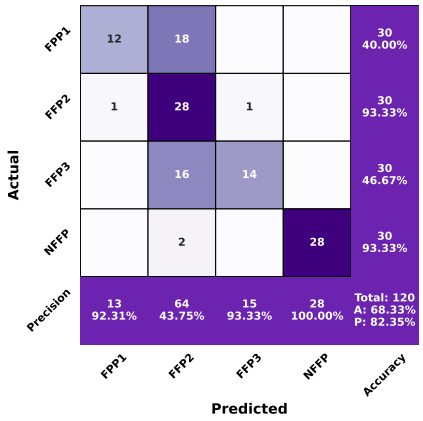

Fig. 3. The confusion matrix obtained by swinv2-large in Protocol 2.

TABLE VII
EVALUATION RESULTS OF BEST OBJECT DETECTION MODELS FINE-TUNED ON PROTOCOL 1 WITH 20 EPOCHS

Model	Accuracy	Precision	Recall	F1	mAP50	mAP50-95
YOLO11n	0.9675	0.9696	0.9655	0.9676	0.9825	0.9295
YOLO11s	**0.9800**	**0.9777**	0.9751	**0.9764**	0.9863	0.9482
YOLO11m	0.9712	0.9686	0.9737	0.9712	0.9847	0.9464
YOLO12n	0.9775	0.9715	0.9773	0.9744	0.9864	0.9305
YOLO12s	0.9712	0.9731	**0.9779**	0.9755	**0.9879**	**0.9489**
YOLO12m	0.9700	0.9648	0.9634	0.9641	0.9800	0.9396

is important to clarify that, unlike image classifiers, which consistently produce a single output for every test image, object detectors can identify a variable number of objects. This means that, depending on the scene, an object detector might label multiple objects, a single object, or even no objects. It may also be cases in which the classes of interest present in the image are not detected (false negatives), as well as those in which objects of interest are detected, but there are none (false positives). YOLO handles these cases via the internal class "Background". Furthermore, the metrics reported in Tables VII and VIII are computed internally by YOLO considering an IoU equal to or greater than 50%. This generates a discrepancy with the Precision in the confusion matrices, computed only on the classes of interest and not considering the IoU. Regarding Protocol 1, Figure 4, depicting the confusion matrix of YOLO12s, reveals that the model effectively differentiates quite well between the different classes of the problem, with smaller accuracies in recognizing Non-Medical (NM) masks. In contrast, Protocol 2 demonstrates its difficulty in managing this case, as the results shown in Figure 5 are evidently inferior to those of Protocol 1. The model appears to have learned to distinguish the FFP2 and NFFP cases effectively, with the notably impressive and somewhat surprising results achieved for the FFP3 class warranting special mention. On the contrary, it does not seem to have succeeded in sufficiently discretizing the FFP1 class from the FFP2.

VII. CONCLUSION

The research concerns recognizing the mask or respirator and assessing their protection level. First, an improved version of the FMR-DB dataset has been presented, with more images

TABLE VIII
EVALUATION RESULTS OF BEST OBJECT DETECTION MODELS FINE-TUNED ON PROTOCOL 2 WITH 20 EPOCHS

Model	Accuracy	Precision	Recall	F1	mAP50	mAP50-95
YOLO11n	0.6167	**0.8151**	0.7795	**0.7969**	0.8262	0.7503
YOLO11s	0.7000	0.6644	**0.8728**	0.7544	0.8503	0.7805
YOLO11m	0.6583	0.7156	0.7995	0.7552	0.8097	0.7550
YOLO12n	**0.7333**	0.7246	0.8454	0.7803	**0.8648**	**0.8093**
YOLO12s	0.6250	0.7581	0.7626	0.7604	0.8336	0.7791
YOLO12m	0.5917	0.6175	0.8353	0.7101	0.7855	0.7327

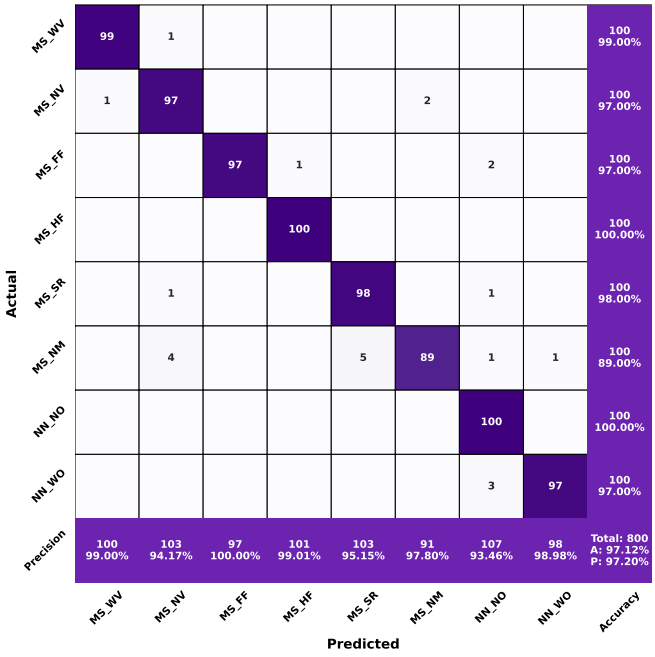

Fig. 4. The confusion matrix obtained by YOLO12s in Protocol 1.

and new labeling to perform the object detection task. Then, the dataset has been used to train different image classification and object detection models in two protocols: Protocol 1 concerns the recognition of the mask or respirator, while Protocol 2 regards the assessment of the protection level that, to the best of our knowledge, no previous research has addressed yet. For Protocol 1, both the image classification and object detection tasks have achieved high values for the considered performance metrics. For Protocol 2, the results were lower. Still, it was interesting to understand better the underlying difficulty of the task, which turns out to be quite challenging even for humans. Future work aims to expand the dataset further to address the problem of evaluating the protection level. It may also be interesting to consider manufacturer-

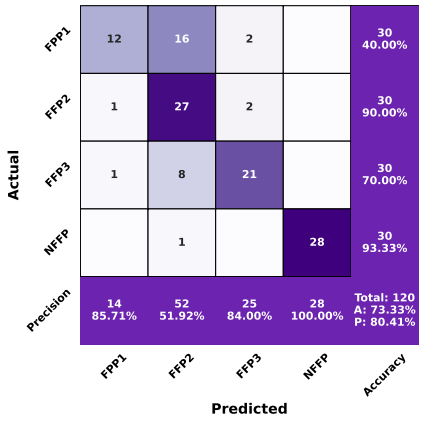

Fig. 5. The confusion matrix obtained by YOLO12n in Protocol 2.

specific solutions for better accuracy, exploiting recurring color schemes or letterings to identify their protection level.

REFERENCES

[1] World Health Organization, "Advice on the use of masks in the context of COVID-19: interim guidance, 5 june 2020," World Health Organization, Technical documents, 2020.

[2] Centers for Disease Control and Prevention, "Types of masks and respirators," CDC Guidelines, 2021, https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/types-of-masks.html, [Online]. Accessed on March 25, 2025.

[3] Occupational Safety and Health Administration, "Respiratory protection standards," OSHA Standards, https://www.osha.gov/respiratory-protection/standards, [Online]. Accessed on March 25, 2025.

[4] Food and Drug Administration, "Face masks, including surgical masks, and respirators for COVID-19," FDA Regulations, 2021, https://www.fda.gov/medical-devices/personal-protective-equipment-infection-control/faqs-face-masks-respirators-and-other-ppe, [Online]. Accessed on March 25, 2025.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," 2016. [Online]. Available: https://arxiv.org/abs/1506.02640

[6] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, p. 303–338, Jun. 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[8] T. Pham, B. Tran, D. Pham, and L. Nguyen, "Classifying between masked faces and normal faces with CNN and SSH," https://github.com/aome510/Mask-Classifier/blob/master/paper/paper.pdf, 8 2019, [Online]. Accessed on March 25, 2025.

[9] Z. Guo, W. Zhou, L. Xiao, X. Hu, Z. Zhang, and Z. Hong, "Occlusion face detection technology based on facial physiology," in *2018 14th International Conference on Computational Intelligence and Security (CIS)*, 2018, pp. 106–109.

[10] S. Yoon and S.-C. Kee, "Detection of partially occluded face using support vector machines," in *IAPR Workshop on Machine Vision Applications (IAPR MVA 2002)*, 12 2002, pp. 546–549.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[12] P. Bhandary, "Mask classifier," https://github.com/prajnasb/observations, 3 2020, [Online]. Accessed on March 25, 2025.

[13] A. Rosebrock, "COVID-19: face mask detector with OpenCV, Keras/TensorFlow, and deep learning," https://www.pyimagesearch.com/2020/05/04/covid-19-face-mask-detector-with-opencv-keras-tensorflow-and-deep-learning/, 5 2020, [Online]. Accessed on March 25, 2025.

[14] A. Muritala, "Mask classifier," https://medium.com/@abimbolamuritala65/covid-19-face-mask-detection-20216652ab6f, 5 2020, [Online]. Accessed on March 25, 2025.

[15] G. Jignesh Chowdary, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Face mask detection using transfer learning of inceptionV3," in *Big Data Analytics*, L. Bellatreche, V. Goyal, H. Fujita, A. Mondal, and P. K. Reddy, Eds. Cham: Springer International Publishing, 2020, pp. 81–90.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] 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. [Online]. Available: https://arxiv.org/abs/1704.04861

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," 2019. [Online]. Available: https://arxiv.org/abs/1801.04381

[19] F. Chollet, "Xception: deep learning with depthwise separable convolutions," 2017. [Online]. Available: https://arxiv.org/abs/1610.02357

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[21] S. E. Snyder and G. Husari, "Thor: a deep learning approach for face mask detection to prevent the COVID-19 pandemic," in *SoutheastCon 2021*, 2021, pp. 1–8.

[22] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "MaskedFace-net – a dataset of correctly/incorrectly masked face images in the context of COVID-19," *Smart Health*, vol. 19, p. 100144, 2021.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, dec 2021.

[24] A. C. Marceddu, R. Ferrero, and B. Montrucchio, "Mask and respirator detection: analysis and potential solutions for a frequently ill-conditioned problem," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2022, pp. 1056–1061.

[25] Humans In The Loop, "Medical mask dataset," https://humansintheloop.org/resources/datasets/medical-mask-dataset/, 2020, [Online]. Accessed on March 25, 2025.

[26] Y. Wang, "Which mask are you wearing? Face mask type detection with TensorFlow and Raspberry Pi," https://medium.com/towards-data-science/which-mask-are-you-wearing-face-mask-type-detection-with-tensorflow-and-raspberry-pi-1c7004641f1, 2020, [Online]. Accessed on March 25, 2025.

[27] A. C. Marceddu and B. Montrucchio, "Recognizing the type of mask or respirator worn through a CNN trained with a novel database," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 1490–1495.

[28] B. Batagelj, P. Peer, V. Štruc, and S. Dobrišek, "How to correctly detect face-masks for COVID-19 from visual information?" *Applied Sciences*, vol. 11, no. 5, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/5/2070

[29] A. C. Marceddu and B. Montrucchio, "Facial masks and respirators database (fmr-db)," in *IEEE Dataport*, 2020.

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

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[32] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[34] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[35] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, 2024.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[37] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[38] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 133–16 142.

[39] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: attention-centric real-time object detectors," 2025. [Online]. Available: https://arxiv.org/abs/2502.12524

[40] G. Jocher and J. Qiu, "Yolov11," Ultralytics, 9 2024, https://docs.ultralytics.com/it/models/yolo11/#overview [Online]. Accessed on March 25, 2025.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101