IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication 00, 0000, date of current version 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.DOI

Unleashing the Transferability Power of Unsupervised Pre-Training for Emotion Recognition in Masked and Unmasked Facial Images

MORENO D'INCÀ[1⋆], CIGDEM BEYAN[1⋆], RADOSLAW NIEWIADOMSKI[2], SIMONE BARATTIN[1], NICU SEBE[1]

[1]Department of Information Engineering and Computer Science (DISI), University of Trento, Trento, Italy
[2]Department of Informatics and Technology, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy

Corresponding author: Cigdem Beyan (e-mail: cigdem.beyan@unitn.it). ⋆ indicates equal contribution.

ABSTRACT Facial expressions are an essential part of nonverbal communication and major indicators of human emotions. Effective automatic Facial Emotion Recognition (FER) systems can facilitate comprehension of an individual's intention, and prospective behaviors in Human-Computer and Human-Robot Interaction. However, FER faces an enduring challenge, commonly encountered in real-life, of partial occlusions caused by objects such as sunglasses and hands. With the onset of the COVID-19 pandemic, facial masks become a major obstruction for FER systems. The utilization of facial masks exacerbates the occlusion issue since these cover a significant portion of a person's face, including the highly informative mouth area from which positive and negative emotions can be differentiated. Conversely, the efficacy of FER is largely contingent upon the supervised learning paradigm, which necessitates costly and laborious data annotation. Our study centers on utilizing the reconstruction capability of a Convolutional Residual Autoencoder to differentiate between positive and negative emotions. The proposed approach employs unsupervised feature learning and takes as inputs facial images of individuals with masks and without masks. Our study puts particular emphasis on the transferability of the proposed approach to different domains in comparison to current state-of-the-art fully supervised methods. The comprehensive experimental evaluation demonstrates the superior transferability of the proposed approach, highlighting the effectiveness of the unsupervised feature learning pipeline. Despite outperforming more complex methods in some scenarios, the proposed approach is characterized by relatively low computational expense. The source code of the proposed approach, along with the facial images created for this study, will be publicly accessible *following the acceptance of this paper*.

INDEX TERMS Facial emotion recognition, facial mask, partial occlusions, affective computing, unsupervised pre-training, human-robot interaction

I. INTRODUCTION

In the last two decades, several researchers proposed models for automatic emotion recognition from nonverbal cues such as voice activity [1], [2], [3], body motions [4], [5], [6], touch [7], as well as their combinations [8], [9]. However, the most often considered indicators of emotional states are facial expressions [10], [11], [12], [13]. In particular, emotion recognition from facial images (referred to as *Facial Emotion Recognition (FER)*) has attracted a tremendous number of researchers [14]. FER is useful to design and develop complex human-machine interfaces [10] for a wide number of applications such as social robotics [15], [16], therapy, diagnosis, and health-care applications [17], virtual training and serious games [18].

In recent years, the FER methods have attracted increasing attention and achieved remarkable performance by integrating deep learning architectures [14], [19], [20]. Still, partial occlusions of the face, e.g., by hands, hairs, sunglasses, scarves, and so forth, are challenges for the FER systems and make them less effective in some cases. Upon the presence

of the COVID-19 pandemic, facial masks have become a major source of partial occlusions, and consequently, several solutions were proposed to perform FER for the facial images with masks [21], [22], [23], [24], [25], [26], [27], [28], [29]. Indeed, using facial masks creates a particularly challenging condition for FER since the masks typically cover half of a person's face, and importantly the mouth area from which highly informative cues for emotion recognition can be extracted [30]. Moreover, in the real world, while some people may wear a facial mask, others may not be able to wear it, or some might use the mask only for a limited period and later take it off. Motivated by such cases, it is important to develop a FER model that can deal with the images of people wearing a mask (referred to as F_m throughout the paper) as well as images of faces without a mask (referred to as F_{um}).

The success of FER predominantly reckons on the supervised learning paradigm in which the data annotation is expensive and laborious. Importantly, obtaining highly reliable emotion labels is tough [8] since the perception of emotional expressions depends on several factors such as gender and culture [31]. There exists a few attempts to perform unsupervised learning: Xiao et al. [32] apply Restricted Boltzmann Machines (RBMs), and Yu et al. [33] use Cycle Generative Adversarial Network (CycleGAN), for this purpose.

This paper tackles the FER problem in images of individuals who may or may not be wearing facial masks. It employs *Unsupervised Feature Learning* (also called *Unsupervised Pre-training*) [34] to address this challenge. The primary advantage of our approach is the elimination of a time-consuming annotation process for feature learning [35], [8]. The proposed method leverages the reconstruction capability of a *Convolutional Residual Autoencoder*. The rationale behind our proposal is to develop a model that can extract informative features applicable to both masked (F_m) and unmasked (F_{um}) facial expression recognition, as well as mask detection, without the need for specific pre-training. Furthermore, our approach is designed to be applicable in real-life scenarios, as it does not strictly require the presence of masks in the target images.

Our model and its application are majorly different from the unsupervised learning-based prior arts [32], [33]. First, none of them [32], [33] uses autoencoders. Furthermore, RBMs applied in [32] have relatively less time efficiency compared to the autoencoders (notice that the computation remains intractable for regular-sized RBMs because its complexity is exponential even in the size of the smallest layer, see [36] for details). On the other hand, the approach in [33] requires neutral facial images (and annotations) for image generation, and suits only small-scale data as mentioned by the authors. Importantly, neither of these works [32], [33] has been tested on F_m images or in-the-wild large-scaled datasets or in case of partial occlusions.

According to a widely employed model of emotions, emotional experiences can be represented in a two-dimensional (2D) space of valence and arousal [37]. In this model, valence determines whether a state is negative (unpleasant) or positive (pleasant), while arousal refers to the degree of activation, ranging from low arousal (deactivated) to high arousal (activated). Representing emotions along dimensions offers several advantages, with the primary benefit being the facilitation of constructing computational models [38]. The 2D model, in particular, is extensively utilized for emotion recognition in domains such as human-computer interaction and human-robot interaction [39]. In these contexts, the knowledge of emotional valence is particularly critical for guiding the progression of interactions, necessitating fast and reliable detection [40]. Motivated by these findings, we utilize our approach to distinguish between *positive* and *negative emotions*. When comes to the human ability to perceive and differentiate positive and negative emotions from the face, several studies confirm that this is related to specific facial areas. For instance, according to [41], positive emotions are mainly perceived by humans from the motion of the lower part of the face. It was also demonstrated that the presence of facial masks decreases humans' ability to perceive and recognize emotions from the face [42], [43], [44]. Such findings make the problem of automatic discrimination of positive and negative emotions from the facial images of people wearing a mask particularly interesting. Therefore, we examine the effect of facial masks on the proposed method's performance such as examining if the positive emotion detection performance of the proposed method is relatively lower in F_m images.

In this paper, we particularly study the *transferability* of FER systems capable of recognizing emotions from F_m and F_{um} images. Unsupervised feature learning has the potential to provide a more robust adaptation to real-world applications due to the fact that it does not require (labeled) re-training when the domain changes [35], [8], [45]. In this regard, we investigate the following cross-dataset scenarios to evaluate whether:

- an unsupervised feature learning-based approach (i.e., Ours) performs better than fully supervised methods (i.e., state-of-the-art (SOTA)) when the pre-training model and classifier share the same domain, but the testing dataset differs.
- an unsupervised feature learning-based method (i.e., Ours) performs better than fully supervised approaches (SOTA) when the pre-training domain differs from the domains used for training and testing the classifiers.

The first case is particularly relevant for evaluating the performance of methods in real-world scenarios where a domain gap often exists between the training data and the deployment domain. This case helps assess the robustness and generalization capabilities of the methods when faced with variations in the testing dataset. On the other hand, the second case assumes that during deployment, a portion of the data becomes available for fine-tuning the pre-trained models, whether they are supervised or unsupervised. This scenario reflects a situation where some adaptation can be performed on the pre-trained models using limited additional

data to enhance their performance in the specific deployment domain. By considering both cases, we can gain insights into the effectiveness of the methods in addressing domain gaps and the potential for further fine-tuning to improve performance in specific deployment scenarios.

The comprehensive experimental analysis conducted demonstrates the superior transferability of our proposed method compared to state-of-the-art (SOTA) approaches in the aforementioned cross-dataset settings (refer to Section IV-D). Notably, our method offers the added advantage of lower computational costs compared to several SOTA methods. This feature has been instrumental in integrating our approach into a social robot as part of the EU Horizon 2020 SPRING project (GA #871245)[1], whose memory is restricted, particularly, when performing several tasks at the same time (Sec. V). Furthermore, when the proposed method was evaluated with the traditional set-up i.e., the pre-training, training, and testing splits are formed from the same dataset, there exist some cases that the proposed method performs better than more complex fully-supervised methods (e.g., having multi-head attention), which is remarkable to spot given our model's significantly lower number of parameters and fewer FLOPs (Sec. IV-C). Notably, the proposed autoencoder is also good at differentiating F_m images from F_{um} images (aka face-mask detection) without requiring additional pre-training different from the one applied for FER (Sec. IV-F). The code of the proposed method and the facial masked images curated within this study will be made publicly available *upon the acceptance of this paper*.

The remainder of this paper is structured as follows. Section II provides an overview of previous research on Facial Expression Recognition (FER) methods capable of recognizing emotions from masked (F_m) images. It also reviews the datasets that have been used for evaluating such methods. The proposed method is introduced in Section III, which outlines the design of the convolutional residual autoencoder, and the inference stage, and includes implementation details. Section IV presents the experimental analysis, including the construction of datasets and the obtained results. In Section V, we describe an application of our method in a real-world scenario, where our model is integrated into a PAL Robotics ARI robot designed to provide assistance in hospital settings. Finally, the paper concludes with a summary and discussions in Section VI.

II. RELATED WORK

With the continuous advancement of deep learning methods, Facial Emotion Recognition (FER) systems have shown remarkable performance improvements in recent times. However, the challenge of face occlusion has emerged as a significant concern due to the increased use of facial masks as a precautionary measure during the COVID-19 pandemic. It's important to note that in some countries, the usage of facial masks remains mandatory, while in others, it may be limited to specific sensitive locations like hospitals. Nevertheless, there are still individuals who voluntarily choose to wear masks, especially in densely populated enclosed spaces. Numerous studies have demonstrated a decrease in the human ability to recognize emotions when a person is wearing a facial mask [42], [43], [44]. These studies confirm that individuals tend to focus primarily on the eyes rather than the mouth for emotion recognition. However, the number of studies that address automated FER in the presence of facial masks remains relatively limited. In Table 1, we provide a summary of such methods while discussing them in detail in the subsequent sections.

A. FER SYSTEMS CAPABLE TO RECOGNIZE EMOTIONS FROM MASKED FACES

To distinguish positive and negative emotions in F_m images, Yang et al. [24] improved effectiveness of MobileNet [46] and VGG19 [47] by fine-tuning them with relevant facial images. That is the first study showing that *i)* FER can feasibly be performed on the images with the generated (simulated) facial masks, *ii)* masked faces decrease the performance of the models trained on F_{um} images, implying that model fine-tuning with F_m images is needed, and *iii)* MobileNet fine-tuned with F_m images performs better than the VGG19 counterpart. However, it is also observable that *iv)* the proposed solutions perform well only when the front view mask is used and *v)* the models are insufficient for small-size training data. The experimental analysis [24] lacks differently shaped or colored masks and does not perform training and/or testing on both F_m and F_{um} images unlike we perform in this work. Barros and Sciutti [21] use FaceChannel [48], which is an adaptation of VGG16 [47] with much fewer parameters, composed of 10 convolutional layers with batch normalization and ReLU, and 4 pooling layers. They tested several pre-training and fine-tuning combinations for the estimation of arousal and valence values. The results show that the pre-training FaceChannel [48] on the original AffectNet dataset [49], and then fine-tuning all layers of the network with masked-AffectNet performs the best no matter the testing data is masked or not. It is also highlighted that supervised pre-training with F_{um} images improves the results, and training the network from scratch with F_m images lowers the performance significantly while it is also not sufficient to only fine-tune the last convolutional layer. The model of [21] was tested on a larger dataset compared to [24], and brought in important findings. However, the experimental analysis was limited to one dataset and one type of mask.

In [22], the authors propose a two-stage deep-attention model to address the face mask problem in FER for three emotions (positive, negative, and neutral). In the first stage, a binary deep model recognizes whether an image contains a mask or not, and generates attention heatmaps to roughly distinguish the masked facial parts from the unobstructed regions. The second stage of the method utilizes the binary attention heatmaps and feature embeddings of the deep model and further includes fully connected layers to perform

[1]https://spring-h2020.eu/

TABLE 1: The summary of SOTA FER systems capable of recognizing emotions from F_m images. The preferable method would be independent to mask detection and face segmentation, tested on multiple large-scale in-the-wild datasets (no role-play, no in-lab. settings), able to process both input types F_m and F_{um} images, being trained on several mask types to potentially better generalize, and able to perform well both on small and large scale datasets. DNN refers to deep neural networks. Sup. and Unsup. stand for supervised and unsupervised, respectively.

Ref.	Feature Learning	DNN	Independent to correct mask detection	Independent to correct face segmentation	Dataset(s)	Inputs: F_m or F_{um}	Mask Types
[24]	Sup.	MobileNet, VGG19	✓	✓	multiple, small-scale in-lab. (private)	F_m	synthesized, real, surgical
[21]	Sup.	FaceChannel	✓	✓	single, large-scale, in-the-wild	F_m	synthesized, surgical
[22]	Sup.	ResNet, VGG w/attention	✗	✗	multiple, large-scale, in-lab. (private), in-the-wild	F_m & F_{um}	synthesized, real, surgical
[25]	Sup.	ResNet w/attention	✓	✓	multiple, large-scale, in-lab., in-the-wild	F_m	synthesized, surgical
[29]	Sup.	Vision Transformer	✗	✗	multiple, large-scale, in-lab. (private), in-the-wild	F_m & F_{um}	synthesized, real, 8 types
[27]	Sup.	MobileNet	✗	✓	single, large-scale in-the-wild	F_m & F_{um}	synthesized, surgical
[23]	Sup.	VGG19, ResNet50 InceptionV3	✗	✗	small-scale frontal, role-play	F_m	synthesized, surgical
[28]	Sup.	CNN	✗	✗	single, small-scale frontal, role-play	F_m	synthesized, surgical
Ours	Unsup.	Autoencoder + MLP	✓	✓	multiple, small/large-scale, in-the-wild	F_m & F_{um}	synthesized, 162 types

FER in the way that the model pays more attention to the unmasked region but less to the masked region. The same authors later on proposed a deep learning pipeline based on face parsing and a vision Transformer with a cross-attention mechanism [29] motivated by the findings of [22], which shows a performance increment upon injecting attention over the mask area. The architecture in [29] consists of three components: *1)* unmasked facial region segmentation using a pre-trained face parsing model, *2)* feature map extractor of pre-trained ResNet50 [50] followed by a multi-layer Transformer encoder, and *3)* fusion of patches from the face mask branch and the feature map patches with the classification token ([CLS]) from the unmasked face branch with a Multilayer Perceptron. The results of [22] remarkably surpass the performances of MobileNet and VGG19 presented in [24] while the performance of [29] is the best out of all. As shown in [29], the computational cost in terms of FLOPs and the number of parameters, the model in [22] is 45 times and five times higher than MobileNet of [24], respectively. Similarly, the model of [24] is 18 times more than MobileNet [24] in terms of FLOPs and seven times more than MobileNet [24] in terms of the number of parameters. As reported in Sec. IV, out of all methods, our proposed method is the most computationally efficient one. Additionally, different from [22], [29], our method excludes the need of detecting the location of the mask and it is able to perform FER in both F_m and F_{um} images within a single model. Instead, [22], [29] requires additional classifiers to perform FER on F_{um} images to compensate for the performance, which would increase the model complexity. Nevertheless, as shown empirically, our unsupervised feature learning stage is able to learn relevant features to be able to distinguish F_m and F_{um} images from each other very effectively. Moreover, it is important to highlight that [29] dependents on a face parsing model requiring to be pre-trained on F_m images, includes two pre-trained ResNet50 [50], a pre-trained transformer, and an MLP head while we rely on a convolutional residual autoencoder, which is pre-trained with an unsupervised manner and an MLP head for classification. The model of [29] was tested on eight types of facial masks, which is the highest number in the literature but still majorly lower than what we tested in this paper.

Another study using a Convolutional Neural Network (CNN) with an attention mechanism is [27]. In that work, the authors additionally check whether the performance of FER system processing F_m images is comparable to humans' performance. Similar to [22], [29], their model [27] requires a mask detector, which in their case is a fine-tuned MobileNet [46]. If a mask is detected, then only the part of the face around the eyes is kept as a Region Of Interest (ROI) and that ROI is classified by a ResNet50 [50] pre-trained on such cropped images. In case of a mask is not detected, then, the entire face is considered as an ROI, which is classified by another ResNet50 [50] to detect discrete emotions (happiness, surprise, anger, sadness, fear, disgust, and neutral). In conclusion, their FER system outperformed humans. It is important to notice that [22], [29], [27] all require additional models to perform FER on F_{um} images and majorly focus on FER on F_m images. However, such a preference can perform poorly, especially in real-world deployment in which the mask detectors fail to detect the existence of a mask or localize the facial masks incorrectly. Therefore, as performed in our study, we claim that a single model able to learn feature representations from both F_m and F_{um} images is beneficial in real-world processing, also promoting less computational

complexity. Shehu et al. [23] likewise compared several pre-trained CNN architectures: VGG19 [47], ResNet50 [50], and InceptionV3 [51] for discrete FER (anger, disgust, fear, happy, neutral, sad and surprise) within four settings of images: *i)* without a mask, *ii)* with a mask covering the lower face, *iii)* a partial mask with a transparent mouth window, and *iv)* with sunglasses. Keeping in mind that the evaluation was performed on a single constraint dataset (in terms of images, which were all captured frontally and the number of instances): extended Cohn-Kanade (CK+) [52], that study [23], in line with [27], shows that the aforementioned models can perform better than humans when the facial area is covered more than 15%. Importantly, human mainly confuses the neutral class with positive/negative emotions, instead, the automated models are able to differentiate the neutral class from emotion classes, but sometimes confuse the negative and positive emotions [23].

To sum up, none of the SOTA has performed unsupervised feature pre-training to develop a FER system capable of recognizing emotions from both F_m and F_{um} images. We also show that our autoencoder supplying feature representations to perform FER can be used for mask detection without the need for any alterations. Instead, in addition to requiring labeled data, the more recent (and more effective) SOTA involves several pre-trained models to detect the mask location, and in some cases to perform FER in F_{um} images. The proposed method's transferability is the best as confirmed by extensive experiments. Among all SOTA, we present one of the most efficient architectures in terms of FLOPS and the number of parameters. It is remarkable that our solution is able to surpass several fully supervised SOTA while performing almost equally well on F_m and F_{um} images.

B. EXPERIMENTAL SETUP OF RELATED WORK

Several earlier approaches in the field have not been thoroughly tested on unconstrained, real-world datasets, as can be noticed in studies such as [23], [28]. Additionally, some of these methods have not consistently been evaluated using publicly available datasets, as seen in research works like [24], [22], [29]. Furthermore, the experimental analysis of these approaches often revolves around a single dataset, limiting the breadth of their evaluation, as evident in papers such as [24], [21], [27], [23], [28]. Some datasets used in the evaluation of [24], [22], [29] were collected in the laboratory environment, composed of a single ethnical group, which might be a concern since a FER system trained on one ethnical group might not generalize well to others given that facial expressions might vary from culture to culture [53]. It is also worth highlighting that there is currently no widely adopted (masked) dataset used for comparing the performance of SOTA methods in this domain. Consequently, we curated our datasets from existing in-the-wild, large-scale FER datasets having valence annotations.

Similar to previous studies [21], [25], we incorporated AffectNet [54], which is widely recognized as a large-scale database for facial expression, valence, and arousal in unconstrained settings, into our evaluation. In addition to AffectNet, we included two other unconstrained FER video datasets, namely Aff-wild2 [55], [56], [57], [58], [59], [60], [61], [62], [63] and AFEW-VA [64], [65], both of which provide valence annotations. Similar to the approaches discussed in the related work, we also employed synthesized masks in our study. We conducted visual inspections to ensure that the facial masks were correctly positioned, and we discarded any images where the masks were improperly placed. It is worth noting that different research groups used different facial mask generators, such as those mentioned in [66], [28], [21]. Furthermore, it is important to highlight that several studies only tested their methods on a single type of mask [48], [23], [27], [24]. However, in our work, we introduced a wide variety of masks for evaluation purposes. While [29] stands out for using eight types of masks, the number of mask variations used in their study is still limited compared to the diverse range of masks we utilized in our research.

Importantly, our experimental setup is different from SOTA since we use both F_m and F_{um} images in training and testing. We claim that such a scenario is more suitable given the current evaluation of COVID-19 such that it is possible to observe both masked and unmasked individuals in our daily life. In terms of experimental analysis, we specifically focus on the cross-dataset performance of our model with respect to SOTA. This transferability has not been investigated by earlier art before, however, we argue that it allows an understanding of the real-world robustness of the methods.

III. PROPOSED METHOD

The proposed Convolutional Autoencoder (AE), visualized in Fig. 1-*top* is composed of an encoder having three main residual blocks, each featuring three convolutions with ReLU and a max-pooling operation. The input image of this network is of dimension $64 \times 64 \times 3$. A single residual block has the 2D-kernels 3×1, 1×3, 3×1. The output of the encoder has a size of 2048. The encoder employs residual connections; particularly the first layer of each block is shared among the block itself and the skip connection, the output of the block is then summed with the output from the skip connection.

The decoder is the transpose version of the encoder employing the same structure that takes as input the latent space from the encoder reconstructing the original image. Each decoder block uses a transpose-convolutional layer with ReLU and batch normalization. A single transpose-convolutional block has the 2D-kernels 3×1, 1×3, 3×1. Decoder has also residual connections and we apply max-unpooling at the beginning of each decoder block while it should be noted that batch normalization is employed only for the decoder.

The reconstruction objective function of our AE is the Mean Squared Error (MSE):

$$\mathcal{L}_{MSE} = \tfrac{1}{2}\mathbb{E}_{\mathbf{X}\sim\mathcal{B}}\big[\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2\big], \tag{1}$$

where $\mathbf{X}$ is the input image, and $\|\cdot\|_F$ denotes the Euclidean norm of the vector obtained after flattening the tensor $\mathbf{X}$. The

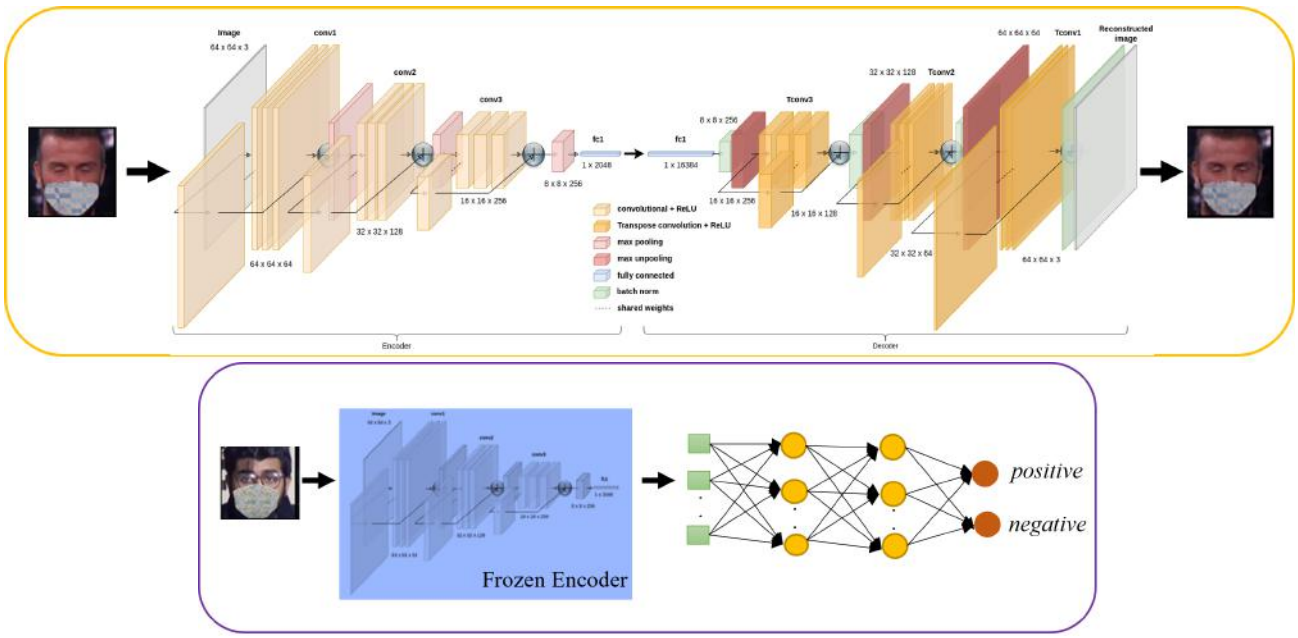

FIGURE 1: Proposed convolutional autoencoder trained with Mean Squared Error loss (top). Downstream task; positive/negative emotion classification learned with an MLP using the features extracted from the frozen encoder of our convolutional autoencoder trained unsupervised way (bottom).

MSE loss in (1) is minimized by using ADAM optimizer over mini-batches $\mathcal{B}$ and the reconstructed data are defined as:

$$\hat{\mathbf{X}} = \mathbf{D}_\theta \circ \mathbf{E}_\varphi(\mathbf{X}), \tag{2}$$

The MSE loss has the learnable parameters θ, φ updated by mini-batch gradient descent, where we estimate

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{B}}\left[\mathcal{L}_{MSE}(\theta,\varphi)\right] = \mathbb{E}_{\mathbf{x}\sim\mathcal{B}}\left[\|\mathbf{x} - \mathbf{D}_\theta(\mathbf{E}_\varphi(\mathbf{x}))\|_F^2\right],$$

by averaging the MSE loss $\mathcal{L}_{MSE}$ over the mini-batch $\mathcal{B}$.

Inference. Once the proposed AE is trained with MSE, without using the labels of the data (aka unsupervised pre-training), following the representation learning literature, we freeze the AE and use it only to extract features for the training/testing data, which are used to train/test a linear classifier (see Fig. 1-bottom).

That linear classifier is a Multilayer Perceptron (MLP) composed of two layers with parametrized ReLu as the activation function, trained to perform the classification of positive and negative emotions. The training of the MLP is performed with Focal Loss [67] motivated by the fact that it could be able to better handle the class imbalance problem, if any.

Implementation Details. The proposed AE was trained for 20 epochs using a combination of Real-World Masked Face (RMFD) and Real-World Masked Face-V2 (RWMFD) datasets [68] and the FER datasets described in Sec. IV-A. RMFD and RWMFD [68] are two datasets containing real (i.e., not synthesized) images of people wearing a mask. We used them in our unsupervised pre-training to allow our AE to learn from real-world F_m images since the FER datasets we used in this study involve synthesized facial masks. Notice that RMFD and RWMFD [68] were used only during the training of AE, but not for training or inference with MLP. Following unsupervised pre-training, the MLP was trained for 45 epochs. In all our experiments, we exploited ADAM as the optimizer with $1e-3$ when the batch size was set to 256. During training, we applied random rotation with a degree range of $[-60, 60]$ to augment the data.

IV. EXPERIMENTAL ANALYSIS

This section first describes the curated FER datasets used for the evaluation of the proposed method and the state-of-the-art models (SOTA) in Sec. IV-A. Then, we introduce the implementation details of the SOTA in Sec. IV-B, which we compare against the proposed method in Sec. IV-C within the same-dataset evaluation setting. Sec. IV-D reports the results of the cross-dataset analysis, which corresponds to the main research questions sought that are with respect to model transferability. Following that, we discuss our model's predictions in terms of positive and negative emotion classes as well as on F_m and F_{um} images in Sec IV-E and highlight our model's capability for mask detection in Sec. IV-F.

A. DATASETS & EVALUATION METRICS

We utilized three large-scale, in-the-wild FER datasets that provide valence annotations. These are: AffectNet [54], Aff-wild2 [55], [56], [57], [58], [59], [60], [61], [62], [63] and AFEW-VA [64], [65]. The AffectNet dataset [54] is one of the largest image-based datasets for FER, including 287651

training, and 4000 validation images annotated manually. We follow the studies in the literature, e.g., [21], [25], [22] using the validation set for model evaluation. The images from the AffectNet dataset exhibit variations in size. The valence annotations provided in the dataset are numerical values ranging between -1 and $+1$. In contrast to AffectNet [54], there is currently no other publicly available large-scale (an important characteristic to train deep models effectively) FER dataset that meets the criteria of being image-based, collected in-the-wild, and providing valence annotations. To overcome the limitations of available image-based FER datasets with valence annotations, we curated our own datasets by extracting images from two existing unconstrained, large-scale video-based FER datasets: Aff-wild2 and AFEW-VA. The Aff-wild2 dataset [55], [56], [57], [58], [59], [60], [61], [62], [63] is composed of 558 videos collected from Youtube including 458 subjects. The valence values are between -1 and $+1$. Lastly, the AFEW-VA dataset [64], [65] contains 600 video clips selected from movies including indoor and outdoor scenes. That dataset provides a wide spectrum of facial expressions, captured in various circumstances with natural head pose movements, complex backgrounds, and severe occlusions [64], [65]. The valence annotations are per frame in a range between -10 to 10.

To ensure consistency in cross-dataset analysis, we discretized the valence annotations in both the AffectNet/Aff-Wild2 and AFEW-VA datasets. We categorized values smaller than zero as belonging to the negative class and values greater than zero as belonging to the positive class. While FER datasets may exhibit variations in the range of valence annotations, the sign of the valence (positive or negative) is crucial for FER analysis, as emphasized in prior work [54].

The F_m images were created from F_{um} images of the original datasets by using the facial mask generation method proposed in [69]. That mask generator [69] provides five different mask types (surgical, N95, KN95, cloth, gas mask), which we used all except the gas mask. It also provides 24 different patterns that can be applied to mask types while it is also allowed to modify the color and intensity of the mask color. To generate the masks, we randomly selected the mask type, pattern, and color for each image in a dataset. We also changed the intensity of the color randomly. This resulted in 162 different facial masks. Since each mask type has multiple templates based on angle, they cover a wide range of face tilts, resulting in accurate F_m images [69]. Still, we applied a manual visual inspection to discard the facial images of having the mask misplaced. Sample F_m images are given in Fig. 2. The datasets' final curation is summarized as follows and the numbers of F_m and F_{um} images in the training and testing splits are given in Table 2.

a) Dataset 1: Its training and testing splits are composed of randomly selected 50% of the original (F_{um}) images of AffectNet combined with the F_m images generated from the other 50% of the dataset. The training and testing instances were kept the same as supplied by the original dataset.

FIGURE 2: Samples of F_m images obtained by applying Anwar and Raychowdhury's method [69] to the original F_{um} images.

b) Dataset 2: Once we removed highly similar faces from the video clips of Aff-Wild2, which refers to faces appearing in consecutive frames with the same emotion type, we utilized facial mask generation techniques described in [69] on the remaining images. The original dataset's provided training and testing splits were retained. We ensured that if a certain type of image (either F_m or F_{um}) appeared in the training set, its corresponding counterpart would not be present in the test set, and vice versa. Moreover, the identities across training and testing splits are not overlapping.

c) Dataset 3: To create an in-the-wild image-based dataset from AFEW-VA, we initially removed highly similar faces. These were defined as faces appearing in consecutive frames of the videos with the same emotion type. Subsequently, we applied the mask generation technique from [69] to the remaining images. Any facial images where the generated mask was inaccurately placed were discarded from the group of F_m images. However, their corresponding original images were retained as F_{um} images. These images, although relatively challenging, are still valuable for evaluation purposes. Rather than completely excluding them from the evaluation, which is the common approach followed by state-of-the-art (SOTA) methods, including them as F_{um} images contribute to a more comprehensive assessment. In this case, it is possible for the same identities to appear in both the training and testing splits, but with differences in head orientation, emotion classes, and image types (F_m or F_{um}).

As observed in Table 2, Datasets 1-3 exhibit a slight imbalance in the number of F_m and F_{um} images within their respective training sets. This may pose an additional challenge for FER models. However, we deliberately avoided manipulating the training splits to achieve balanced classes, as imbalanced data is a common occurrence in real-world applications [70]. We believe that the dataset curation undertaken in this study is a valuable contribution, particularly considering the absence of a standardized benchmark. The effectiveness of the proposed method and SOTA are measured with F1-score ($F1$).

B. THE STATE-OF-THE-ART METHODS

We adopted several fully supervised SOTA methods in order to compare their efficiency and effectiveness against the proposed approach. Each of them was first pre-trained for the mask detection task using the relevant real-world, large-

TABLE 2: Details of the datasets used in the experimental analysis.

Dataset	Source	# of Training Images			# of Testing Images		
		Unmasked	Masked	Total	Unmasked	Masked	Total
1	AffectNet	143825	130205	274030	1999	1903	3902
2	Aff-Wild2	145920	105032	250952	31873	24624	56497
3	AFEW-VA	5658	7024	12682	631	781	1412

TABLE 3: Evaluation of the proposed method and the SOTA on Datasets 1, 2, and 3 in terms of F1-score. The best results are indicated in **bold** and the second best results are given underlined. The symbol ↑ implies that a higher value is preferred.

Method	Feature Learning	F1 (↑)		
		Dataset 1	Dataset 2	Dataset 3
Barros & Sciutti [21]	supervised	48.8	26.9	75.2
ResNet50 [50]	supervised	66.2	41.2	79.2
(Proposed) Know. Dist.	supervised	70.3	44.1	83.8
ViT [71]	supervised	38.3	29.9	58.2
Swin-L [72]	supervised	46.3	44.9	56.8
ViT (w/ResNet50) [71], [73]	supervised	**71.0**	**65.7**	87.7
Proposed	unsupervised	58.8	46.6	**95.4**

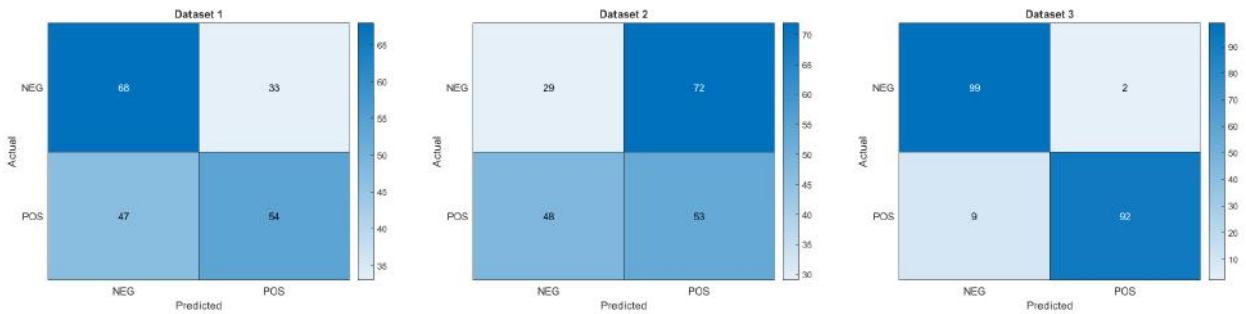

FIGURE 3: Confusion matrices correspond to the proposed method trained and tested on Dataset 1 (left), Dataset 2 (middle), and Dataset 3 (right).

TABLE 4: Computational cost of the SOTA and the proposed method in terms of floating-point operations per second (FLOPs) and the number of parameters. ⋆ represents the results based on our implementation. The number of parameters (# Params) of the knowledge distillation methodology is given in terms of the sum of the teacher and the student, respectively, while the FLOPs are given for the student model since it is the one used in deployment time. The FLOPs of the proposed method include the autoencoder and the MLP. In parenthesis, we state the MLP's FLOPs. The values are indicated in **bold** and the second best values are given underlined. The symbol ↓ indicates that a lower value is considered better.

Method	Computational Cost	
	FLOPs (↓)	# Params (↓)
Barros & Sciutti [21]	0.1G⋆	**2.1M**⋆
ResNet50 [50]	25.7G	4.3M
(Proposed) Know. Dist.	0.1G	24.3M+1.5M
ViT [71]	17.6G	86.7M
Swin-L [72]	34.5G	197M
ViT (w/ResNet50) [71], [73]	17.6G	86.7M
Proposed	0.45G (**0.09G**)	2.75M

scale, in-the-wild datasets: RMFD and RWMFD [68] in line with the proposed autoencoder's pre-training. The further implementation details are described as follows.

The methodology of Barros & Sciutti [21]. We utilized the FaceChannel network [48] in our implementation, adapting its last layer to suit the binary classification task (i.e., softmax). Following the implementation details described in [21], we employed the same search space for the number of layers and units per layer. The ADAM optimizer [74] was utilized, with a learning rate of $1e-3$, and the model was trained using cross-entropy loss.

Proposed Knowledge Distillation. One can observe that several SOTA methods, e.g., [24], [29] adapted MobileNet while some other studies, e.g., [23] showed the effectiveness of InceptionV3 to recognize emotions from masked faces. Eyiokur et al. [75] recently showed SOTA results of InceptionV3 and MobileNet for face mask detection in unconstrained environments. Motivated by such studies, we developed a fully supervised method performing knowledge distillation between InceptionV3 and MobileNet. Knowledge

distillation among neural networks is wildly used for several applications when resource constraints are in place [76]. This is also valid for us since our final aim is to integrate the FER approach into a Social Robot. However, knowledge distillation has not been tested before by a relevant prior study. Given the final model's efficiency by being computationally less expensive (i.e., MobileNetV3 has a lower number of parameters and fewer floating-point operations per second (FLOPs), see Table 3 for details) compared to several SOTA, and its on-par performance against several CNN-based SOTA, we argue that the proposed knowledge distillation model is the fairest fully supervised counterpart of the proposed unsupervised feature learning-based method.

Knowledge distillation revolves around the utilization of two neural networks, namely a student and a teacher. The teacher network is designed to be larger, prioritizing high classification performance without considering resource limitations. On the other hand, the student network is a smaller network specifically designed to meet low resource requirements while aiming to achieve similar results to the teacher network. Our teacher, InceptionV3, was pre-trained on ImageNet [77], while our student network, MobileNetV3, was trained from scratch. The PyThorch implementations of these networks were exploited by keeping their architectures unchanged except for the classification layers which were updated to fit our aim. The teacher network was fine-tuned using the cross-entropy loss function. The student's objective function was the weighted sum of a soft and a hard loss. The soft loss has the goal to distill the knowledge of the teacher via soft targets computed over the predictions of the teacher (i.e., soft outputs). The soft loss function is the Kullback–Leibler divergence of the soft outputs and soft targets. The hard loss is the cross-entropy function between the student predictions and the ground truth. Consequently, the overall student objective function is: $L_s = L_{soft} \cdot T^2 \cdot w_s + L_{hard} \cdot (1 - w_s)$ where w_s is the soft weight set to 0.9, L_{soft} is the soft loss and L_{hard} is the hard loss. It is important to multiply the soft loss by T^2 (T stands for the temperature value) when using both hard and soft targets. This ensures that the relative contributions of the hard and soft targets remain roughly unchanged if the temperature used for distillation is changed while experimenting with meta-parameters [76]. For all corresponding experiments, ADAM [74] was used as the optimizer with a $1e-3$ learning rate in line with the proposed unsupervised feature learning-based method. The input images were resized to several scales such as 64×64 or 224×224 (note that we did not observe significant performance differences for the same dataset experiments). The best results were obtained when the number of epochs and the batch size were taken as 20 and 64, respectively, for the InceptionV3 while MobileNetV3 was trained up to 30 epochs with 64 batch size.

ResNet50. Another network that has been frequently adapted by SOTA is ResNet50 [27], [29], [25], [23]. Moreover, since the autoencoder of the proposed method employs residual connections (which as shown through the ResNet family, is an important characteristic to avoid the problem of vanishing gradient and mitigating the degradation problem, resulting in poor learning for deep networks), we included that model into the comparative study. We used the ResNet50 pre-trained on ImageNet and employed softmax loss at the end of the network. For training, we used $1e-3$ and $1e-4$ learning rates and 0.0005 weight decay parameters. The optimization was performed with ADAM [74] while the inputs were scaled to 256×256. Training of the models was executed with batch sizes from 64 to 256.

Vision Transformer (ViT) and Swin Transformer (Swin-L). Since the attention mechanisms have been adopted by several FER studies (see Sec. II for details), we also included the Visual Transformers [71] and Swin Transformer [72] into our comparisons following the implementation details given in [78], [24].

It is currently very well-known that the performance of ViT majorly drops when it is trained from scratch compared to being fine-tuned. The reason for this is its limited feature extraction capacity appearing in case of not using the guidance of large-scale datasets. In other words, ViT has less induction bias compared to a CNN, thus, it is harder to train, and large-scale datasets help to compensate for the performance gap [73]. A typical way to handle this situation is to use a pre-trained CNN. In [78], it was shown that the performance of ViT trained from scratch can be 38% less than using pre-trained CNN together with ViT for FER. In this study, we integrated a ResNet50 model (i.e., the pre-trained model on ImageNet was further fine-tuned on RMFD and RWMFD datasets [68] as mentioned at the beginning of this section) to extract features from the last convolutional layer of it, referring to the FER model in [78] as well as the FER model for F_m images in [24]. The number of layers and the head were set to 4 and 6, respectively for the multi-layer transformer encoder in line with [24]. The hidden dimension of the MLP head was set to 1000 [24] or 3072 [78]. The learning rate was initialized as $5e-3$ [24], [78], following a warmup of 250 steps and a cosine learning rate decay [24], [78]. The model was trained up to 300 epochs with ADAM optimizer [74] with cross-entropy loss when the batch size was varied between 64 to 256.

In the case of the Swin-L Transformer [72], we fine-tuned it up to 50 epochs using a learning rate of $1e-2$ with the ADAM optimizer.

C. COMPARISONS AGAINST THE STATE-OF-THE-ART

Even though our main focus is to study the *transferability* of the proposed method with respect to other approaches, we first report a comparative study across our model and the prior art on the same-dataset setup to draw us an empirically validated comparative method out of all SOTA (see Table 3 and Fig. 3).

The results highlight the better performance of ViT [71] used together with pre-trained ResNet50, on average. How-

TABLE 5: Cross-dataset analysis when the testing dataset is different from the pre-training and training datasets. The best results of each metric are given in bold. Notice that the pre-training of the proposed method is unsupervised, i.e., without using the labels. The symbol ↑ implies that a higher value is preferred.

Method	Feature Learning	Pre-training Dataset	Classifier		F1 (↑)
			Training Dataset	Testing Dataset	
Know. Dist.	supervised	-	1	2	38.3
Proposed	unsupervised	1	1	2	**44.7**
Know. Dist.	supervised	-	2	1	44.0
Proposed	unsupervised	2	2	1	**58.4**
Know. Dist.	supervised	-	1	3	**60.4**
Proposed	unsupervised	1	1	3	53.2
Know. Dist.	supervised	-	2	3	46.8
Proposed	unsupervised	2	2	3	**51.2**

ever, our approach surpasses ViT with ResNet50 when tested on datasets whose scalability is relatively smaller such as the case of Dataset 3. For relatively larger datasets such as Dataset 2, our model demonstrates the second-best performance after ViT with ResNet50 by surpassing all other fully supervised methods. Without using pre-trained ResNet50, ViT [71] underperforms in all datasets. The proposed Knowledge Distillation approach, overall, achieves better results compared to Barros and Sciutti [21] and ResNet50 even though its student component is much lightweight compared to both approaches. Based on the confusion matrices, it is evident that the proposed method exhibits a higher detection rate for negative emotions compared to positive emotions in Dataset 1. Conversely, for Dataset 2, the positive emotion detection rate surpasses the negative emotion detection rate. In the case of Dataset 3, the detection rates for each class are relatively closer to each other, although negative emotions still tend to be better predicted.

In terms of computational cost (see Table 4) and performance trade-off, the best-performing fully supervised method is the proposed knowledge distillation model. Therefore, we use that model to compare its performance against the proposed method within the cross-dataset evaluations performed to validate the transferability.

D. CROSS-DATASET ANALYSIS

The cross-dataset analysis includes two types of investigation. In the first one, we evaluate the models' performances when the datasets used in the pre-training and during the training of the classifier are the same, but the classifier's testing dataset is different. Such experiments are relevant given that there is often a domain gap between the training/validation data and the testing domain in real-world applications. The corresponding results are given in Table 5.

Table 5 shows that the majority of the time the proposed unsupervised feature learning-based model's transferability is superior to the proposed fully supervised knowledge distillation model. The only exception occurred when Dataset 1 was used as the training dataset and the testing is performed on Dataset 3. Still, even in the further case, the performance gap between the two models is lower than the former, i.e., the proposed unsupervised feature learning-based model surpasses the knowledge distillation. Overall, a drop in performance is possible due to the domain gap between the datasets. Particularly, training on either Dataset 1 or Dataset 2 significantly decreases the performance on Dataset 3 compared to both training and testing on Dataset 3.

The second type of cross-dataset analysis is to evaluate the models' performances when the pre-training dataset is different from the dataset the classifiers are trained and tested on. Such a setting simulates real-world applications in which one typically has models trained on one dataset (so-called pre-trained models) but further needs to be fine-tuned on another dataset whose distribution is the same as the testing dataset but different from the pre-training dataset. We evaluated the performance of the knowledge distillation model in two settings:

(a) The teacher model was trained on the pre-training dataset, and then the student network was trained on the same dataset. Furthermore, the student network was fine-tuned with the classifier's training dataset and tested with the classifier's test set. All layers of the student network were fine-tuned.

(b) The teacher network was trained on the pre-training dataset, and then the student model was trained on the same dataset. Consequently, the student network was fine-tuned with the classifier's training dataset and tested with the classifier's test set. Only the last layer of the student was fine-tuned.

The corresponding results are given in Table 6. Herein, we used Dataset 1 and Dataset 2 in pre-training, and Dataset 3 was used for the classifier's training and testing. It is a common practice that model pre-training is performed on relatively larger datasets. In this vein, we did not perform pre-training on Dataset 3 given that it is the smallest dataset out of all (otherwise it is highly likely that a catastrophic forgetting would happen, therefore the transferability cannot be studied). Also in such cases, the proposed unsupervised feature

TABLE 6: Cross-dataset analysis when the pre-training dataset is different from the dataset the classifier is fine-tuned and tested on. The best results of each metric are given in bold. Notice that the pre-training of the proposed method is unsupervised, i.e., without using the labels. See the text for the description of (a) and (b). The symbol ↑ implies that a higher value is preferred.

Method	Feature Learning	Pre-training Dataset	Classifier		F1 (↑)
			Training Dataset	Testing Dataset	
Know. Dist. (a)	supervised	1	3	3	81.8
Know. Dist. (b)	supervised	1	3	3	69.5
Proposed	unsupervised	1	3	3	**95.8**
Know. Dist. (a)	supervised	2	3	3	84.7
Know. Dist. (b)	supervised	2	3	3	60.2
Proposed	unsupervised	2	3	3	**94.5**

learning-based approach surpasses the proposed knowledge distillation model for both settings (a) and (b), once again proving its better transferability. It is notable that pre-training on Dataset 1 slightly improves the results (from 95.4% to 95.8%) of our unsupervised feature learning-based method with respect to the one given in Table 3 (i.e., the same-dataset analysis) and pre-training on Dataset 2 improves the results of proposed knowledge distillation with respect to the same-dataset analysis (from 83.8% to 84.7%).

E. FURTHER ANALYSIS ON EMOTION CLASSES

Table 7 reports the F1 score of the proposed method for positive and negative emotion classes as well as its performance on F_m and F_{um} images. Overall, the proposed method is better at performing emotion classification on F_{um} images than the F_m images. The average performances on F_{um} images and F_m images do not have a significant gap.

Specifically, the performance of the proposed method on positive emotions is better than the detection of negative emotions when the test dataset is Dataset 2. This finding is on par with the model in [27] showing that positive emotions are better identified since they are more related to the eye regions while negative emotions are mainly related to the mouth region and therefore barely recognized. However, when the test datasets are Dataset 1 or Dataset 3, we observe the opposite results such that the negative emotions are detected better than the positive emotions.

In general, these last two results obtained for our model are consistent with the results on the identification of positive/negative emotions from F_m images by humans, which are also unequivocal [43], [44].

F. MASK DETECTION

Given that several SOTA assume the input image includes a facial mask to perform FER (e.g., [24], [23], [28]) or first detect the mask location (e.g., [22], [29], [27]) and then apply the corresponding FER network, we believe that it is interesting to show the mask detection capacity of our method. To do so, instead of training an MLP for emotion classification, we train an MLP for a mask detection task, importantly by keeping the unsupervised feature learning stage unchanged across emotion recognition and mask detection tasks. We obtained 99.8%, 99.9%, and 99.4% F1 scores for Datasets 1, 2, and 3, respectively.

V. CASE STUDY: HUMAN-ROBOT INTERACTION

Within the EU Horizon 2020 SPRING project (GA #871245) the proposed unsupervised feature learning-based method was integrated into the robot ARI developed by PAL Robotics. The project focuses on Socially Assistive Robots (SARs) and their applications in healthcare.

In one scenario, where robots are envisioned to serve as receptionists and interact with humans, their role would involve greeting human agents, gathering basic information about the purpose of their visit, and guiding them to the appropriate specialist. Given that a single robot would interact with numerous individuals, it is crucial for the embedded models to demonstrate robustness in dealing with a wide variety of facial images. Additionally, while mask requirements have been relaxed in many countries, there remains a significant number of individuals who continue to wear masks as a precaution against respiratory diseases. Hence, it becomes essential to handle both masked (F_m) and unmasked (F_{um}) facial images equally well, ensuring optimal performance across both categories.

For this scenario, it is essential that the robot is able to detect the faces and then recognize as soon as possible (i.e., nearly in real-time) the emotional states of a human interaction partner, or at least the emotional valence (i.e., whether the person feels a negative or positive emotion). This basic information about the emotional state of the human agent can be crucial for the course of the interaction. It is so because there exists a risk that the person might withdraw from interacting with the robot when being in a negative emotional state (see [79] for more information) or feeling not at ease (e.g., she/he might feel offended or misunderstood by a robot). In such cases, the robot should be able to detect the negative valence and potentially modify its behavior to maintain the interaction.

Fig. 4 shows a simulation of this use case. The robot is interacting with a human agent wearing a mask. In this demo, on the touchscreen, the images captured by the head

TABLE 7: Positive/negative emotion discrimination of the proposed method together with its emotion classification performance (in terms of F1 score) on the F_m and F_{um} images. P, T, and Te stand for pre-training, training, and testing, respectively. D1, D2, and D3 mean Datasets 1, 2, and 3.

	P(D1)T(D1)Te(D1)	P(D2)T(D2)Te(D2)	P(D3)T(D3)Te(D3)	
Positive	48.0	**61.5**	93.8	
Negative	**69.5**	31.7	**97.1**	
Masked	60.0	**46.1**	93.9	
Unmasked	**63.7**	44.6	**96.2**	
	P(D1)T(D1)Te(D2)	P(D2)T(D2)Te(D1)	P(D1)T(D1)Te(D3)	P(D2)T(D2)Te(D3)
Positive	**55.4**	43.5	32.9	41.2
Negative	34.0	**73.2**	**73.6**	**61.1**
Masked	55.4	56.8	**48.9**	**52.2**
Unmasked	**55.6**	**60.0**	42.4	44.9
	P(D1)T(D3)Te(D3)	P(D2)T(D3)Te(D3)		
Positive	94.3	92.6		
Negative	**97.3**	**96.4**		
Masked	94.3	92.9		
Unmasked	**96.5**	**96.5**		

FIGURE 4: An example of interaction between an ARI robot and a human agent. On the touchscreen, the images captured by the head camera of the robot as well as the FER result per image are displayed.

camera of the robot as well as the detected emotion label are displayed.

VI. DISCUSSION AND CONCLUSIONS

We have presented a method exploiting the reconstruction capability of a Convolutional Residual Autoencoder to differentiate between positive and negative emotions when F_m and F_{um} images are the inputs. This method performs unsupervised feature training, therefore, learns the relevant latent features without using labeled data, which brings an advantage since gathering relevant data annotations for emotion recognition could be challenging. The detailed experimental analysis demonstrates the better transferability of the proposed method, which is an important property for its real-world application. It is also important to highlight that our method has lower computational costs compared to several existing solutions, allowing us to integrate it into a social robot that performs several tasks simultaneously. When the proposed method was evaluated within the same dataset setting, its better performance compared to more complex methods such as Vision Transformers is noticeable particularly when the quantity of data is relatively scarce (e.g., around 13K training images). The proposed autoencoder is also good at differentiating F_m images from F_{um} images, thus, performs mask detection without a need for additional pre-training (i.e., pre-training different from the one applied for emotion recognition) while recognizing emotions mostly equally well in both F_m and F_{um} images.

To summarize, the contributions of this study are:

- The presented unsupervised pre-training leverages the reconstruction property of autoencoders. The classifier trained on top of the learned features is lightweight in terms of the number of parameters and the FLOPs.
- Motivated by the SOTA, we propose an additional FER method, which is fully supervised, and based on knowledge distillation. That method is the best among all prior SOTA given the performance and efficiency tradeoff.
- The experimental analysis of the two proposed methods (a) unsupervised feature learning-based and (b) knowledge distillation-based show the better generalizability of the unsupervised one.
- The proposed autoencoder, without any additional need of training, can be effectively used for other downstream

tasks than FER such as mask detection.
- Overall, the proposed unsupervised feature learning-based method performs equally well on F_m and F_{um} images.
- The datasets used in this study, containing the F_m images generated from existing in-the-wild, large-scale FER datasets (F_{um}) having the valence annotations, will be shared with the community to serve as a benchmark to foster the following research.
- The effective performance and efficiency of the proposed unsupervised learning-based method allowed us to integrate it into a social robot.

Future work will adapt continual learning strategies and focus on not only positive and negative classes but also the classification of several discrete emotion classes. Another future objective of us is to investigate the social acceptance of a robot equipped with an automatic emotion recognition capability, utilizing the proposed methodology.

ACKNOWLEDGMENTS

This work was supported by the European Union H2020 SPRING project (GA #871245). The authors thank Francesco Tonini, Mouez Khelifi, and Alessandro Conti for their help to integrate the proposed method into the ARI robot.

REFERENCES

[1] A. Lausen and K. Hammerschmidt, "Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–17, 2020.

[2] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody based emotion recognition for mexi," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 1138–1144.

[3] R. K. Sreenivasa, G. Shashidhar *et al.*, "Robust emotion recognition using spectral and prosodic features," in *Springer Briefs in Speech Technology*, 2013.

[4] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, and R. Niewiadomski, "Modeling multiple temporal scales of full-body movements for emotion classification," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[5] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 90–101, 2016.

[6] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 1, pp. 1–31, 2016.

[7] R. Niewiadomski, C. Beyan, and A. Sciutti, "Affect recognition in hand-object interaction using object-sensed tactile and kinematic data," *IEEE Transactions on Haptics*, vol. 16, no. 1, pp. 112–117, 2023.

[8] R. Franceschini, E. Fini, C. Beyan, A. Conti, F. Arrigoni, and E. Ricci, "Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss," in *Proc. of ICRP*, 2022.

[9] Y. Song, Y. Cai, and L. Tan, "Video-audio emotion recognition based on feature fusion deep learning method," in *IEEE Int. Midwest Symposium on Circuits and Systems (MWSCAS)*, 2021, pp. 611–616.

[10] M. Bentoumi, M. Daoud, M. Benaouali, and A. Taleb Ahmed, "Improvement of emotion recognition from facial images using deep learning and early stopping cross validation," *Multimedia Tools and applications*, pp. 1–31, 2022.

[11] T. Hinz, P. Barros, and S. Wermter, "The effects of regularization on learning facial expressions with convolutional neural networks," in *International Conference on Artificial Neural Networks*. Springer, 2016, pp. 80–87.

[12] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE access*, vol. 7, pp. 41 273–41 285, 2019.

[13] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based cnn for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, 2020.

[14] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[15] V. Dimitrievska and N. Ackovska, "Behavior models of emotion-featured robots: A survey," *Journal of Intelligent & Robotic Systems*, vol. 100, no. 3, pp. 1031–1053, 2020.

[16] S. Iengo, A. Origlia, M. Staffa, and A. Finzi, "Attentional and emotional regulation in human-robot interaction," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012, pp. 1135–1140.

[17] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 147–152.

[18] E. Marchi, B. Schuller, A. Baird, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, T. Baltrušaitis, A. Adams, M. Mahmoud, O. Golan, S. Fridenson-Hayo, S. Tal, S. Newman, N. Meir-Goren, A. Camurri, S. Piana, S. Bölte, M. Sezgin, N. Alyuz, A. Rynkiewicz, and A. Baranger, "The asc-inclusion perceptual serious gaming platform for autistic children," *IEEE Transactions on Games*, vol. 11, no. 4, pp. 328–339, 2019.

[19] T. Zhang, "Facial expression recognition based on deep learning: a survey," in *International conference on intelligent and interactive systems and applications*. Springer, 2017, pp. 345–352.

[20] P. Barra, L. De Maio, and S. Barra, "Emotion recognition by web-shaped model," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 321–11 336, 2023.

[21] P. Barros and A. Sciutti, "I only have eyes for you: The impact of masks on convolutional-based facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1226–1231.

[22] B. Yang, W. Jianming, and G. Hattori, "Face mask aware robust facial expression recognition during the covid-19 pandemic," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 240–244.

[23] H. A. Shehu, W. N. Browne, and H. Eisenbarth, "A comparison of humans and machine learning classifiers categorizing emotion from faces with different coverings," *Applied Soft Computing*, vol. 130, p. 109701, 2022.

[24] B. Yang, J. Wu, and G. Hattori, "Facial expression recognition with the advent of face masks," in *19th International Conference on Mobile and Ubiquitous Multimedia*, 2020, pp. 335–337.

[25] D. Gera and S. Balasubramanian, "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition," *Pattern Recognition Letters*, vol. 145, pp. 58–66, 2021.

[26] Z. Yang, K. Nayan, Z. Fan, and H. Cao, "Multimodal emotion recognition with surgical and fabric masks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4678–4682.

[27] G. Castellano, B. De Carolis, and N. Macchiarulo, "Automatic facial emotion recognition at the covid-19 pandemic time," *Multimedia Tools and Applications*, pp. 1–19, 2022.

[28] S. Tegani and T. Abdelmoutia, "Using covid-19 masks dataset to implement deep convolutional neural networks for facial emotion recognition," in *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*. IEEE, 2021, pp. 1–5.

[29] B. Yang, J. Wu, K. Ikeda, G. Hattori, M. Sugano, Y. Iwasawa, and Y. Matsuo, "Face-mask-aware facial expression recognition based on face parsing and vision transformer," *Pattern Recognition Letters*, vol. 164, pp. 173–182, 2022.

[30] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. how individual face parts contribute to successful emotion recognition," *PloS one*, vol. 12, no. 5, p. e0177239, 2017.

[31] X. Fang, G. A. van Kleef, and D. A. Sauter, "Revisiting cultural differences in emotion perception between easterners and westerners: Chinese perceivers are accurate, but see additional non-intended emotions in negative facial expressions," *Journal of Experimental Social Psychology*, vol. 82, pp. 152–159, 2019.

[32] Y. Xiao, D. Wang, and L. Hou, "Unsupervised emotion recognition algorithm based on improved deep belief model in combination with probabilistic linear discriminant analysis," *Personal Ubiquitous Comput.*, vol. 23, no. 3–4, p. 553–562, jul 2019.

[33] Y. Yu, Y. Sun, and Z. Yang, "An unsupervised facial expression recognition method based on cyclegan," in *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, 2022, pp. 669–674.

[34] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 201–208.

[35] G. Paoletti, J. Cavazza, C. Beyan, and A. Del Bue, "Unsupervised Human Action Recognition with Skeletal Graph Laplacian and Self-Supervised Viewpoints Invariance," in *The 32nd British Machine Vision Conference (BMVC)*, 2021.

[36] A. Fischer and C. Igel, "An introduction to restricted boltzmann machines," in *Iberoamerican congress on pattern recognition*. Springer, 2012, pp. 14–36.

[37] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[38] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[39] A. Hong, N. Lunscher, T. Hu, Y. Tsuboi, X. Zhang, S. Franco dos Reis Alves, G. Nejat, and B. Benhabib, "A multimodal emotional human–robot interaction architecture for social robots engaged in bidirectional communication," *IEEE Transactions on Cybernetics*, vol. 51, no. 12, pp. 5954–5968, 2021.

[40] D. Kulic and E. A. Croft, "Affective state estimation for human–robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[41] J. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of personality and social psychology*, vol. 37, pp. 2049–58, 12 1979.

[42] C.-C. Carbon, "Wearing face masks strongly confuses counterparts in reading emotions," *Frontiers in psychology*, vol. 11, p. 566886, 2020.

[43] F. Grundmann, K. Epstude, and S. Scheibe, "Face masks reduce emotion-recognition accuracy and perceived closeness," *Plos one*, vol. 16, no. 4, p. e0249792, 2021.

[44] C. A. Levitan, I. Rusk, D. Jonas-Delson, H. Lou, L. Kuzniar, G. Davidson, and A. Sherman, "Mask wearing affects emotion perception," *i-Perception*, vol. 13, no. 3, p. 20416695221107391, 2022, pMID: 35782826.

[45] G. Paoletti, C. Beyan, and A. Del Bue, "Graph laplacian-improved convolutional residual autoencoder for unsupervised human action and emotion recognition," *IEEE Access*, vol. 10, pp. 131 128–131 143, 2022.

[46] 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," *arXiv preprint arXiv:1704.04861*, 2017.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[48] P. Barros, N. Churamani, and A. Sciutti, "The facechannel: a fast and furious deep neural network for facial expression recognition," *SN Computer Science*, vol. 1, no. 6, pp. 1–10, 2020.

[49] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

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

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[52] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

[53] R. E. Jack, R. Caldara, and P. G. Schyns, "Internal representations reveal cultural diversity in expectations of facial expressions of emotion." *Journal of Experimental Psychology: General*, vol. 141, no. 1, p. 19, 2012.

[54] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *CoRR*, vol. abs/1708.03985, 2017.

[55] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," *arXiv preprint arXiv:2202.10659*, 2022.

[56] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pp. 794–800.

[57] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for heterogeneous multi-task learning: a large-scale face study," *arXiv preprint arXiv:2105.03790*, 2021.

[58] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

[59] ——, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.

[60] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019.

[61] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[62] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1980–1987.

[63] D. Kollias and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3652–3660.

[64] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[65] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 03, pp. 34–41, 2012.

[66] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[67] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *CoRR*, vol. abs/1708.02002, 2017.

[68] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei *et al.*, "Masked face recognition dataset and application," *arXiv preprint arXiv:2003.09093*, 2020.

[69] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," 2020.

[70] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653–1672, 2015.

[71] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[72] 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.

[73] N. Park and S. Kim, "How do vision transformers work?" in *ICLR*, 2022.

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[75] F. I. Eyiokur, H. K. Ekenel, and A. Waibel, "Unconstrained face mask and face-hand interaction datasets: building a computer vision system to help prevent the transmission of covid-19," *Signal, Image and Video Processing*, pp. 1–8, 2022.

[76] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[77] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[78] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, 2021.
[79] M. E. L. Redondo, A. Sciutti, S. Incao, F. Rea, and R. Niewiadomski, "Can robots impact human comfortability during a live interview?" in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 186–189.

MORENO D'INCÀ received his MS in Artificial Intelligence from the University of Trento after completing his Master's thesis at the Queen Mary University of London. He is currently a Ph.D. student at the University of Trento and PicsArt with research interests in image synthesis and human behavior understanding.

CIGDEM BEYAN received her Ph.D. degree in Informatics from the University of Edinburgh, U.K., in 2015. She is currently an Assistant Professor at the University of Trento in the Department of Information Engineering and Computer Science. She has co-authored over 50 papers published in peer-reviewed journals and international conferences. Among her main research interest, there are human behavior understanding, social signal processing, and multimodal data analysis. She is a reviewer of several journals including various IEEE Transactions, and top-tier IEEE/ACM conferences. She is on the Editorial Board of ICES Journal of Marine Science, a Guest Editor in the International Journal of Social Robotics, and has been a Guest Editor in Frontiers in Robotics and AI. She is a member of ELLIS.

RADOSLAW NIEWIADOMSKI received the Ph.D. degree in Computer Science from the University of Perugia (Italy). He is currently an Assistant Professor at the University of Genoa. His research interests include emotion recognition, non-verbal behavior synthesis, and multimodal interaction. He has been involved in several EU research projects, e.g., FP6 CALLAS, FP7 ILHAIRE, and H2020 DANCE. He co-authored over 100 peer-reviewed conference and journal papers.

SIMONE BARATTIN Simone Barattin has completed his Master's degree in Artificial Intelligence Systems at the University of Trento doing his Master's thesis at the Queen Mary University of London. His research focuses on the use of deep learning in computer vision applications such as face anonymization and emotion recognition.

NICU SEBE (Senior Member, IEEE) is a Professor at the University of Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding. He is a fellow of IAPR. He was the General Co-Chair of the IEEE FG 2008 and ACM Multimedia 2013. He was the Program Chair of ACM Multimedia 2011 and 2007, ECCV 2016, ICCV 2017, and ICPR 2020. He is the General Chair of ACM Multimedia 2022 and the Program Chair of ECCV 2024.