

RESEARCH ARTICLE

Across the Universe: Biasing Facial Representations Toward Non-Universal Emotions With the Face-STN

PABLO BARROS^{1,2} AND ALESSANDRA SCIUTTI², (Member, IEEE)

¹Sony, Sens.AI, Research and Development Center, Brussels, Belgium

²Contact Unit, Italian Institute of Technology, 16163 Genova, Italy

Corresponding author: Pablo Barros (pablovin@gmail.com)

The work by Alessandra Sciutti has been supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. G.A. No 804388, wHiSPER.

ABSTRACT Facial expression recognition, as part of an affective computing system, usually relies on solid performance metrics to be successful. These metrics depend significantly on the affective context in which one evaluates this system. While presenting excellent performance on the dataset it was trained on, a facial expression recognition model might drastically fail when one assesses it in a different scenario. Such performance reduction occurs because most facial perception models rely on an extreme generalization concept, focusing on a universal emotion perception system. With the recent findings on the non-universality of emotional perception, generalization of facial encoders seems not to be the optimal path to take. Therefore, exploiting transfer learning toward adapting specific facial features to specific scenarios could address this problem. This paper proposes and investigates a Spatial Transformer Plugin (STN) to rearrange different facial encoders towards particular affective representations from different scenarios. We experiment with our model in eight different facial expression recognition datasets (AffectNet and the derived MaskedAffectNet, OMG-Emotion, FERPlus, ElderReact, EmoReact, FABO and JAFFE datasets) and obtain competitive performance with much less training effort than state-of-the-art models. Besides performance alone, we introduce the STN as a mechanism towards a non-universal emotional perception system and discuss how it rearranges learned perception features to address some specific characteristics of each investigated dataset.


INDEX TERMS Affective computing, facial expression recognition, neural networks, transfer learning.

I. INTRODUCTION

One of the key factors in understanding the lack of adaptability in current automatic facial expression recognition systems comes from the categorization of affect itself. For a long time, the concept of a universal understanding of emotions [26] guided the development of facial expression recognition (FER) systems. The notion that any person in the world can identify one out of six basic emotions independently of their cultural background made the task of labelling and categorizing facial expressions easier [72]. This led to a plethora of artificial systems trained and validated over these predefined concepts. Even affect categorization specializations, such as the popular dimensional arousal and valence

models [46], continue to rely on generalizing affect to claim great performance on emotion expression recognition. This has become even more evident in a recent publication by Coen *et al.* [22] where researchers analysed the presence of predefined affective expressions over millions of YouTube videos from all over the world. However, as discussed by Lisa Feldman [29], Coen *et al.* trained and evaluated their automatic perception model based on a set of predefined and unchangeable emotional concepts, leading their neural network to classify what it was trained to do: sixteen known emotional expressions. The ability of such models to recognize affect from any given scenario is therefore restricted by similar scenarios with which the neural network was trained.

The problem of adaptability is more evident when one deploys these models in real-world scenarios, such as the recent applications in social robots [69]. Usually, their

The associate editor coordinating the review of this manuscript and approving it for publication was Kah Phooi (Jasmine) Seng .

evaluation in cross-dataset experiments (which simulate their application in different scenarios), tends to decrease the FER performance drastically if it is not followed by a computationally heavy fine-tuning or readaptation routine [58]. In most cases, the cause of this lack of adaptability is usually mentioned as different input characteristics and pre-processing, or label distribution. In reality, the problem could lie in the task itself. Recent findings on affect categorization show that emotion perception might not be as universal as we have been led to believe [39], [40], [45].

These recent investigations discuss how interpreting affect comes directly from our world understanding; each person, based on one's expertise, has a certain way of expressing and recognizing affect [35]. In other words, every person sees and understands affect differently, and we adapt and converge towards a known interpretation while interacting. Each carries our affective perception world in this regard: unique and constantly updating.

Translating this view into affective computing, specifically in the development of heavily supervised learning models, we hypothesize that the general understanding of how a facial expression can be categorized is represented by the labelling procedure. Whoever chooses the labels of a dataset is giving to that specific contextual scenario (i.e., all the data samples that compose the dataset) a unique understanding of affect. One must interpret training and evaluating a model in such a dataset that can achieve amazing performances by considering it within the dataset's own constrained characteristics, in particular by considering the labelling decision.

In this study, we address the problem of adapting facial expression recognition by proposing a spatial transformer network (Face-STN) plugin layer that is trained to specify high-level affective features of given facial encoders. Different from traditional fine-tuning and transfer learning, our proposed model is a light-weight neural layer that can be trained with very little effort to improve facial expression recognition. Our Face-STN leverages from visual transformers capability to learn specific characteristics of visual representations [31], [60], but focused on learning the specific affective characteristics of each evaluated scenario.

We evaluate our Face-STN model on specifying the feature representations learned by three different encoders: the strongly supervised FaceChannel [6], the semi-supervised PK, a Generative Adversarial Network [7] that learns facial representations based on self-supervised image reconstruction, and a Contrastive Predictive Coding Network [65] that learns representations by contrasting the specific facial characteristics present on images from the same affective category.

To help us represent different scenarios and therefore evaluate our model in different affective scenarios, we use 8 different datasets in all our experiments. First, we run a baseline scenario for transfer learning and fine-tuning, training the encoders on the one million images drawn from the internet presence on the AffectNet dataset [61]. We then evaluate how our model compares with traditional transfer learning

in seven different scenarios: The monologues and individualized expressions from the OMG-Emotion dataset [5], the internet-crawled images labelled with an affective distribution on the FERPlus dataset [10], the Elder React dataset [56] with recordings of elderly persons, the EmoReact dataset [64] with facial expressions from children, the images from FABO dataset which are composed of acted facial expressions [33], and the Japanese-woman-only expressions from the JAFFE dataset [54]. To investigate a constrained interaction scenario with partial facial expressions, we recently presented a version of the AffectNet dataset where all the images have facial masks added, calling it MaskedAffectNet [9].

To provide a complete understanding of the impact of our Face-STN plugin, we also run a feature analysis that compares the learned representations of each encoder in different transfer learning scenarios and when trained with the plugin. We discuss how our solution compares to existing state-of-the-art results in terms of FER performance for each of these datasets, and how our approach leverages the non-universal affective perception theory to provide a competitive FER solution in most of the evaluated scenarios. Our results show that different from what is believed, training these models with more data would achieve better feature recombination and decision boundary for each specific task. Doing so would not lead to better feature representation, which directly impacts stagnation of the performance observed when applying full-network fine-tuning. Our dataset-specific models achieve competitive performance when compared with complex models, validating our investigation on biasing facial features towards specific tasks. We proceed with an in-depth discussion, through facial features visualization, on how the differences in deep learned facial expressions rely mostly on the dataset chosen to train and evaluate the model, representing a specific task. We conclude by connecting our observations with the non-universal expressions theory, by exemplifying the impact that each affective scenario has on learning emotional representations, and that to have an artificial system able to deal with different scenarios, it is necessary to have fast readaptation on a decision-making level.

Our paper is organized as follows: in the next section (Section II), we introduce our related work and situate the reader about the facial expression adaptability problem; In Section III, we introduce the encoders and detail their implementation, the Face-STN model and all its training and updating mechanisms are proposed; Section IV describe our evaluation and experimental effort, followed by Section V, which exhibits all our results. We discuss our findings in Section VI, and finally conclude the paper in Section VII.

II. RELATED WORK AND IMPORTANT ARGUMENTS

To separate facial representation from affect understanding is somehow intuitive, and in fact has been addressed by several solutions in the past [13], [24]. Most of these have taken this path due to technological limitations. Once convolutional neural networks became a universal solution for

data representation, researchers began giving most of their attention to end-to-end learning of facial expressions [38]. In this section, we present our view on how end-to-end affective perception goes against the concept of non-universality of emotion representation, and thus, presents a limitation on automatic facial expression recognition. We also consider how the current solutions that claim adaptability and transfer learning do not address the problem properly – they provide a shallow layer as a solution instead of dealing with the root of the problem.

A. THE GENERAL END-TO-END AFFECTIVE PERCEPTION

Most of the current state-of-the-art solutions for automatic facial expression recognition (FER) claim to have addressed the problem of global FER by approaching maximum generalization [8], [49], [63]. The majority of these approaches deploy the computational power of artificial neural networks, boosted by data-driven deep learning of faces. The modus operandi of these solutions is to use millions of examples to tune these networks to extract specific facial features that represent and categorize affect. Unfortunately, the learned facial features are biased owing to the very specific scenarios represented by the datasets on which these models are trained and validated. In most of these models, the learned features are comparable with existing human-made modelling such as the Facial Action Units [21], [25], [27]. Coupled with a case-specific good performance, these Units are being perceived as good candidates for a general facial representation system.

The problem these models face when deployed in different or socially constrained scenarios appears when combining these representations into affective categories [51]. Most of these models, mostly for commodity and data availability, categorize affect using standard representations, whether by means of a strict set of categories or dimensional pleasure/arousal/dominance scales. They are not only sensitive to representing only the facial features that are present on the training data, but they are also sensitive to categorizing such features based on given affective labels. Such labels are usually obtained using a transitive bias of giving instructions based on constrained options: the already specified set of categories, or the predefined boundaries for dimensional scales. The generalization aspects of the trained model are bounded to the capability of the labelling procedures.

B. THE RELATION BETWEEN FACES AND CONVOLUTIONAL-BASED RECOGNITION MODELS

If one is to minimize the bias from a pretrained model for affective categorization by providing a computationally light and effective adaptation mechanism, focusing on the representation of facial structures, we could increase the adaptability of affective recognition models in different scenarios. Faces change little. The position of eyes, mouth, and cheeks will be always relatively close to each other [78]. Their representations, different from an affective category, are universal [28], [78]. A healthy person will detect



FIGURE 1. Facial features are differently exposed when expressing affect while using a mask.

different facial structures [75] even on non-facial images [42], leading them to be easy to identify and adapt. This is the case for most facially constrained interactions, like when participants use facial masks. Most current convolution-based emotion expression recognition solutions (the most common ones) already rely on a general facial representation [51], even if implicitly learned by strongly supervised end-to-end learning. Once we can present a soft separation between these facial representations and the affective categorization, it would be much easier to recombine their meaning into a unique world understanding of affective category.

C. THE ARGUMENT OF ADAPTABILITY

When deployed in scenarios that are different from the ones for which models were tuned, most recent affective perception models present difficulty to perform and even to adapt, given that deep neural networks are known to require extreme resources and data-hungry [80]. These models are thus extremely biased towards their application, and they are most often difficult to adapt to specific scenarios [73]. Models without a popular interest and those that do not provide large amounts of available or labeled data are underrepresented world views. One of these scenarios, now in strong evidence given the COVID-19 pandemic, is when social interactions are constrained using personal protective equipment such as facial masks. As most of these neural networks learn how to recognize affect based on a collection of facial features, when some of these features are absent, which is the case when using a mask (illustrated by Figure 1), these models tend to fail [2]. This effect is also observable, albeit on a smaller scale, in humans. However, due to our capability of changing the way we recognize emotions when seem a partially covered face [57], [76], we learn to compensate much better than any deep learning system.

D. THE CURRENT PROBLEMS ON PERSONALIZING AFFECT

The models that get closer to the concept of a strong separation between facial representation and affective

understanding are the ones that claim to provide personalized perception. In these models, the affective concept usually is specialized to a single individual, or a group of individuals that share the same contextual background [67]. Such models generally rely on strong feature representation and on mechanisms to specify features away from the initial affective estimation [70]. Most current solutions focus more on an auditory representation of affect [18], [19], [47]. Although it may be easy to assume that this happens due to the availability of personal auditory information, the reality is other: convolutional neural networks have become experts on image representation, while still struggling to represent auditory features, and specifically speech [53]. Most convnets that deal with speech are extremely complex, and not easily accessible without access to very specific and powerful hardware [1]. Representing speech, and auditory signalling in general, therefore typically occurs with traditional feature extractors that hinder an end-to-end learning approach and facilitate a strong separation between signal representation and affect categorization.

When applied to facial expression recognition, the few models that approach personalization focus on over-specifying the learned features to unique persons [7], [20]. Recently, facial expression representation was attempted to be separated from affective understanding [4], but the proposed model relied on the unique world view of a specific dataset to accomplish both feature representation and affective understanding. Adapting it towards a universal feature representation would demand retraining the entire neural network, and thus, adaptability and representation transfer is not feasible

III. BIASING FACIAL EXPRESSION REPRESENTATION WITH THE FACE-STN

There exist many facial representation models, most of which are based on convolutional neural networks (ConvNets). Hierarchical representation of a ConvNet allows the representation of facial features to emerge within the network layers [14], [62]. The typical facial representation learned by these networks resembles human-made Facial Action Units, which measure different muscle movements to describe a facial expression [43], [55].

Most of these models rely on explicit supervision, coming in the form of a given label, to learn feature maps that represent faces. This process specifies the features towards that specific unique world, represented by the datasets and associated labels that the model is trained with. Other solutions focus on learning facial representations through implicit supervision, such as in the case of convolutional auto-encoders [68], [79], and most recently Generative Adversarial Networks [15], [52], [77].

Most of these solutions bias facial expression representation models towards a unique affective world representation both in the data distribution and presentation, as well as in the labelling process. In this way, most of

these models remain very difficult to adapt towards a novel scenario.

To perform a complete analysis of facial representation, we investigate how different learning schemes contribute to emerging facial representations. In this regard, we investigate the FaceChannel [6], a ConvNet trained with explicit labels; the Prior-Knowledge Generative Adversarial Network (GAN), part of the P-AffMemory model [7], that learns facial representations by identifying real and generated faces; and a novel facial representation based on a Contrastive Predictive Coding network [65], that learns to represent faces based on reconstructing latent representation space itself. Each of these models implements convolutional layers to highlight facial features, and we are interested in investigating the similarities of such features and how we may reuse them on different affective worlds representation.

A. THE IMPACT OF EXPLICIT LABELS WITH THE FaceChannel

The FaceChannel is a recently proposed convolutional neural network with a light-weighted architecture that implements inhibitory layers to improve facial expression representation. It has a total of 2 million parameters, allowing it to be trained from the scratch while making it easily adaptable to other tasks. Our implementation of the FaceChannel has 10 convolutional layers. The last of them is represented by a shunting inhibitory layer [30] and 4 pooling layers. An inhibitory neuron S_{nc}^{xy} , present at position (x,y) of the n^{th} receptive field in the c^{th} layer is defined as:

$$S_{nc}^{xy} = \frac{u_{nc}^{xy}}{a_{nc} + I_{nc}^{xy}} \quad (1)$$

where u_{nc}^{xy} is the activation function of the convolution unit, in our case ReLu, and I_{nc}^{xy} is the activation of the inhibitory units. The passive decay term a_{nc} is also updated during training and is shared among each inhibitory filter.

When training, after the convolutional layers, the FaceChannel implements a fully connected hidden layer implementing a ReLu activation function. This layer is followed by an output layer that implements the direct label decision of the network. This involves a set of neurons implementing a SoftMax activation for categorical classification, or linear activation for a continuous and dimensional representation of affect.

The convolutional layers of the FaceChannel demonstrate the capability of learning different facial representations based on the dataset with which it was trained [6]. The changes are modulated directly from the output layer; the facial representation reflects the labelling distribution of the dataset with which the network is trained. In our investigations, we are interested in understanding the strength of this modulation, and how different the learned feature representation is when training this model with faces collected from different scenarios. Figure 2 illustrates the facial representation layers of the FaceChannel.

B. THE IMPACT OF A RECONSTRUCTION ERROR WITH THE PRIOR-KNOWLEDGE GAN

The Prior-Knowledge (PK) is an autoencoder that learns facial representations by applying an adversarial training routine between real and generated faces. It implements a controllable term that allows the change of affective characteristics, in terms of continuous arousal and valence, to the decoded faces. Besides the encoder and decoder/generator architecture (E and G respectively), the PK also implements three discriminators: the arousal/valence enforcer (D_{em}), the discriminator that guides the latent representation to follow a uniform distribution (D_{prior}), and the adversarial discriminator that ensures the decoded image contains the desired affective information (D_{real}). The PK receives an image (x) and a continuous arousal/valence label (y) as an input and produces a facial latent representation (z) as well as an edited image expressing the chosen arousal/valence (x_{gen}).

The encoder architecture (E) is implemented as four convolutional layers and one fully connected output layer structure. The decoder (G) implements the same structure, though inverted. We do not apply pooling; we use a strided convolution with an order of 2 to provide a dimensional reduction and reduce the network's number of trainable parameters. The encoder represents an RGB image into a latent representation (z), then feeds it concatenated with the desired affective label (y) to the decoder. The entire autoencoder is trained with an image reconstruction loss (L_{rec}) using mean squared error.

The affective information discriminator (D_{em}) guides the encoder to learn facial representations. Recent experiments show that, without this discriminator, the network learned facial representations that did not carry affective content like hair and eye colour [7]. It is implemented as two fully connected hidden layers followed by two linear neurons: one for arousal and one for valence. It is trained using a mean-squared error loss (L_{em}).

The uniform distribution discriminator (D_{prior}) enforces the latent facial representation (z) to be uniformly distributed. It showed to be important to increase the generalization of the model, and to help on the imposition of the affective features within the latent representation [7]. It implements four fully-connected layers, and it is trained using an adversarial loss ($\min_E \max_{D_z} L_{prior}$) between the original distribution of z and an artificial uniform distribution $p_{prior}(z)$.

Finally, the last discriminator (D_{real}) imposes the photorealistic characteristics and enforces that the affective labels (y) are present on the generated images. It implements four convolutional layers, which receive the generated image (x_{gen}), followed by two fully connected layers. Each of the convolutional layers also receives the desired affective information (y), to enforce that it is present on the generated image. This discriminator is trained using an adversarial loss that implements a mean-squared error on the original image and the generated one ($\min_G \max_{D_{img}} L_{img}$).

Previous experiments with the PK demonstrated that generated images carried affective information, but did not

maintain the personal identity [7]. To solve this, we implemented a identity-preserving loss ($\min_{E,G} L_{iden}$) on the reconstructed image. This loss is computed between the original image and the generated one by using the mean-squared error from the last layer representations from a pretrained VGG face [16] encoder.

As is typical for GANS, the PK is quite a sensitive model to be trained, and the impact of each of these losses was defined based on a grid-search focused on minimizing a total loss:

$$\min_{E,G} \max_{D_z, D_{img}} L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{em} + \lambda_3 L_{iden} + \lambda_4 L_z + \lambda_5 L_{img} \quad (2)$$

The coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 served as a balance between each discriminator. Figure 3 illustrates the final architecture of the PK with all the parameters.

C. THE IMPACT OF LATENT REPRESENTATION PREDICTION WITH CPC

Contrastive Predictive Coding (CPC) [65] is a recent self-supervised model that learns to predict the entangled representations of sequences of input stimuli using autoregression. It applies a contrastive InfoNCE loss [65] to enforce data representation which maximizes the reconstruction of future stimuli. For that, it uses an encoder (E) to learn the representation of an image (i) from a sequence (T) of observed stimuli (x_i) and outputs a latent state ($z_x = E(x_i)$), and an autoregressive neural network (A) that integrates a sequence of latent representations ($w \leq T$) into a temporally-contextual latent representation ($C_w = A(z_w)$). This context representation is then used to predict the next element on the sequence (x_k).

Differently from traditional generative models, which focus on learning a representation that is useful to generate or reconstruct the original stimuli, CPC focuses on encoding information that is present on the data sequence. For that, it predicts future stimuli by modeling a log-bilinear model of a density ratio (f_k) between the perceived stimuli sequence (x_w) and the contextual latent representation (C_w):

$$f_k(x_i + x_k, C_w) = \exp(z_T^{i+k} W_{kC_i}) \quad (3)$$

where W_{kC_i} is a linear transformation used for the prediction of the next element in the sequence. The entire network is trained to optimize (f_k) by distinguishing the density ratio between positive and negative samples. Thus, the CPC model learns in a self-supervised manner, estimating the labels directly from the density ratio. As we are interested on learning affective information from the facial expressions, we create positive and negative examples directly from the training data distribution, by clustering samples from the same categorical, or dimensional, representation into positive samples.

We implement our encoder as a series of convolutional layers, and the autoregressor as a GRU network. The entire architecture, with the detailed parameters, is illustrated in Figure 4.

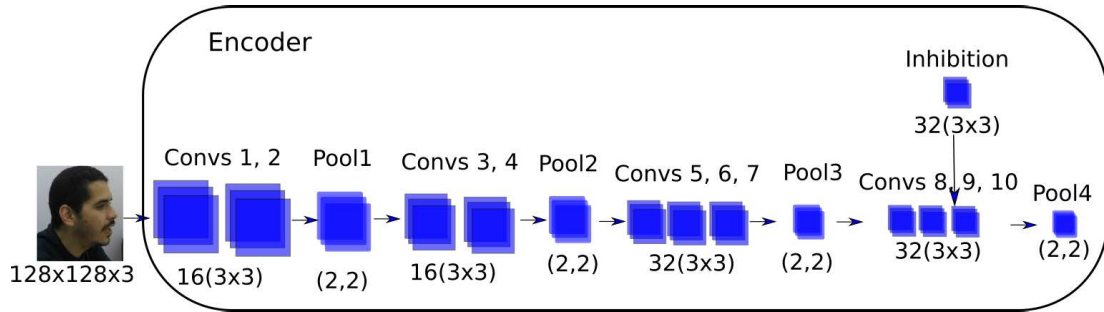


FIGURE 2. FaceChannel architecture used to learn facial expression representations using an explicit supervision.

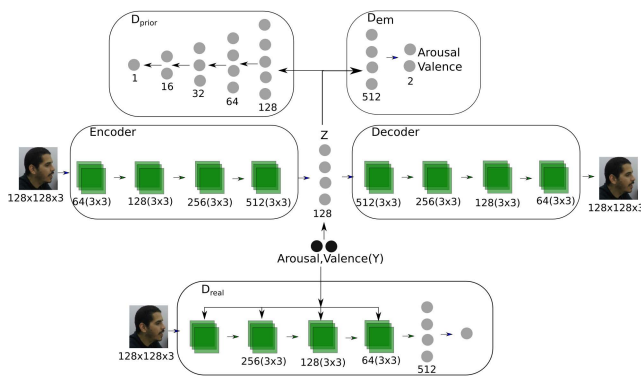


FIGURE 3. PK network with all four discriminators used to enforce a high-fidelity image editing.

Because the learning of the representations in a CPC network is made directly on the latent representations themselves, it does not require many training epochs, neither many examples, as demonstrated by its recent applications in the representation of phonemes [36] and EEG signals [3].

D. THE FACE-STN

One of the most common strategies when adapting facial expression recognition towards different scenarios is to retrain, entirely or partially, a neural network such as any of the three models we introduced in the last section. This enforces that the affective information, from both facial representation and emotional categorization, is somehow depicted by both the convolutional channels and the decision-making layers. The problem when readapting this network towards a novel scenario, is that both the facial representation and the decision-making layers carry an inductive bias from the dataset the model was originally trained with. So, if the new scenario carries any similarity with the originally trained dataset, the tendency is that the emotion recognition performance improves; however, when the scenario is very different, a new and expensive training scheme is usually necessary to achieve a good performance. In doing so, we also change the entire affective representation present on the network and make it less probable to deal with other scenarios.

Spatial Transformer Networks (STNs) have recently been used to learn specific facial characteristics that help on

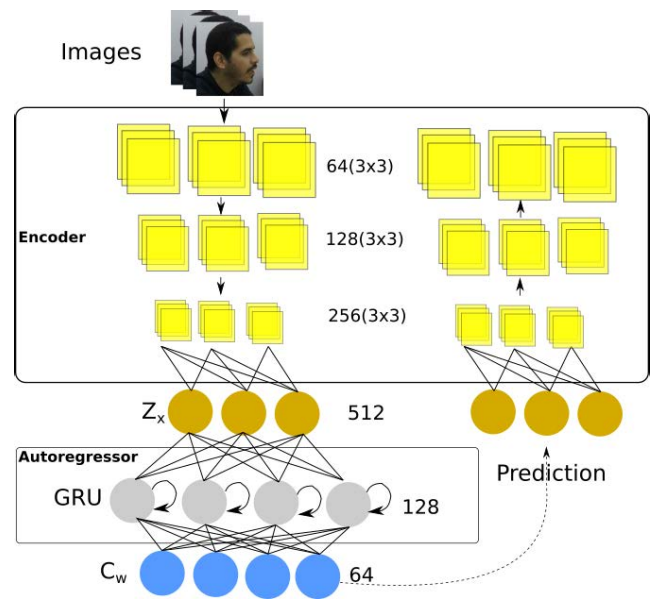


FIGURE 4. Architecture of the Contrastive Predictive Coding (CPC) network used in our experiments.

emotion expression recognition on specific datasets [31], [60]. An STN relies on a localization convolutional neural network that learns specific image transformation parameters, biased by the strongly supervised learning. For faces, making an STN learn different affine transformations can help it to identify specific geometrical characteristics of faces, which might be unique for each dataset.

Our Face-STN is composed of a set of convolutional layers, usually referred to as localization network (L), and receives as input feature maps (F_i) from the convolutional encoder. Often STNs process the input image, but as faces do not change much, and the convolutional channels of the encoders are known to depict facial information, having it applied directly to the feature maps allows us to recombine the learned facial features to deal with the characteristics of a specific dataset. Also, this allows the training of the Face-STN with less data, as it has to learn useful transformations based on already processed input stimuli, and not on the raw image. The Face-STN has as a role to learn the parameters T_θ to perform the affine transformations on the feature maps. After the set

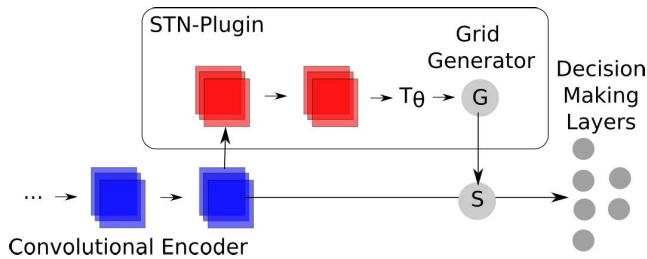


FIGURE 5. Proposed architecture of the Spatial Transformer Network (Face-STN) Plugin.

of convolutional layers of the Face-STN, a grid generator (G) is used, to apply the transformations into different patches of the feature maps. A bilinear sampling kernel [41] (S) sampler is used to select transformations from the grid generator and use them as an additional input to the decision-making of the encoder. Our Face-STN is then trained in a supervised manner, together with the decision-making of each of the encoders.

As the Face-STN is applicable to any convolutional connection to the encoder network, we conducted an exploratory experiment to identify where we want to apply it. The results of this experiment show us that using the feature maps from the last channel of each encoder as input to the Face-STN allowed the best ratio between the number of training parameters and the performance of the network. As such, we illustrate the final Face-STN architecture in Figure 5.

IV. EVALUATING THE LEARNED FACIAL REPRESENTATIONS AND ADAPTATION MECHANISMS

In our evaluations, we want first to investigate the role of traditional fine-tuning and transfer learning mechanisms on learning affective information from faces. Second, we want to evaluate the impact of the Face-STN on biasing the deep visual representations towards affective information. And finally, we want to contrast all these approaches, in objective performance terms, but also on visualizing learned representations.

We have divided our experiments into three settings: first, we run a baseline study to establish the best architectural design of each proposed encoder. We do this by training, evaluating, and fine-tuning them using the AffectNet dataset. Our second experimental setting consists of investigating the capability of each facial encoder to represent the unique characteristics of each dataset and to evaluate if the learned representations can be transferred from one affective to another, with traditional transfer learning and fine-tuning methods. For that, we contrast four training routines: first, we train the entire encoder and the decision-making layer (All layers). Second, we train the last-convolutional layer of the encoder and the decision-making layer (Last Conv-Layer). Third, we train only the decision-making layer (Decision-Making) and fourth we train the entire network from the scratch (Scratch).

We then attach the Face-STN plugin to each encoder, and train them, together with the decision-layer, with all the datasets. This way we can compare the impact of the Face-STN alone with all the other fine-tuning and transfer learning routines. Besides our baseline investigations, we also compare our performance results with existing state-of-the-art models for each dataset. This allows us to evaluate the overall performance of our proposed model, and its impact on the field of facial expression recognition.

For each setting, we propose and explain a series of experiments in the following writeup. We also present specific metrics for each dataset. Each of the used datasets has a unique characteristic, either regarding the image selection and processing, or the affect representation, or both.

A. UNIQUE AFFECTIVE DATASETS

Each of the datasets we use in our experiments (illustrated in Figure 6) has specific characteristics which include image selection and processing, labelling strategy, and data distribution. We also derive a unique decision-making layer for each, illustrated in Figure 13. We individually optimized these as described in our Appendix A. The decision-making layers are connected to the encoder of each model to provide the best performance. The sessions below present each dataset, their unique characteristics, and information about how they were evaluated.

AffectNet [61] is our main baseline and comparison point. It has over 1 million images drawn from the internet, with half of them manually annotated using mechanical Turk. Each image has a single label based on a continuous arousal and valence value. It provides a specific training and validation subset, and we use the concordance correlation coefficient (CCC) [48] for arousal and valence between the models' predictions and the true labels as a performance metric. The images of the AffectNet are centred and are provided as cropped faces. This enforces the encoders to learn facial representations from a large variance of faces, but with a very predictable facial structure, which together with good data distribution, contributes to it mostly be used to train facial expression encoders for other tasks [51]. The labels, although crawled from the internet, were collected based on given concepts of arousal and valence, and thus follow a very specific rule, which makes it possible to be used for benchmarking automatic facial expression recognition models. A simple decision-making layer, composed of fully connected units, and two linear output heads, one for arousal and one for valence, provided the best results for in our exploratory experiments.

The FER+ [10] dataset contains around 31,000 grey-scaled face images crawled from the internet. Each image has a small resolution, of 48×48 , and has a centred and cropped face. To label the images, a crowd-sourced strategy was used, where each labeller was given one out of seven affective categories to choose from: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The authors obtained 10 labellers per image and provide the final label as a distribution of the

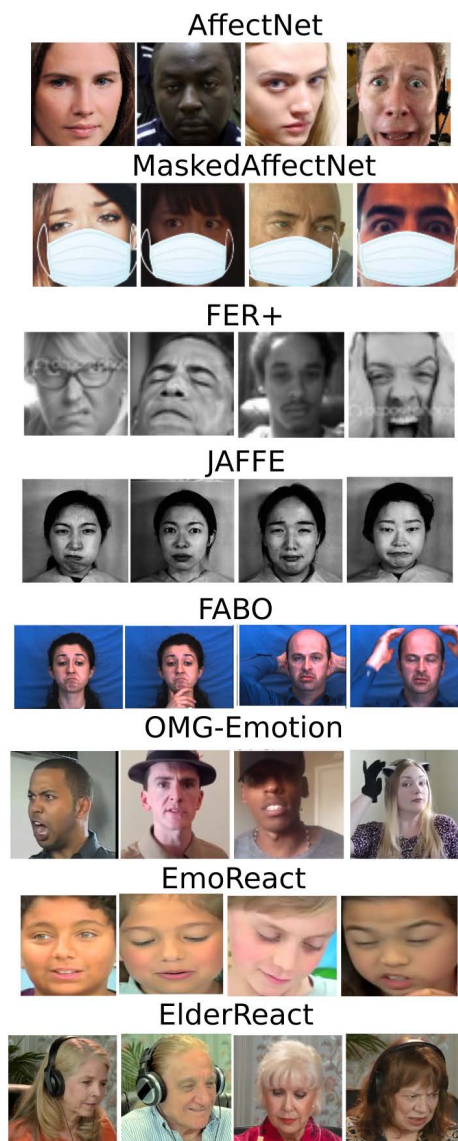


FIGURE 6. Examples of all the datasets used in our experiments.

10 votes. This means that each image is labeled using a composition of the given 7 concepts. The decision-making layer for the FER+ is also based on a fully connected layer but followed by a single SoftMax output. We use the provided train, test, and validation separations in our experiments, and use the accuracy over all the classes as our main performance metric.

The JAFFE [54] dataset contains 213 images from 10 Japanese women performing facial expressions. Each person was asked to perform 3 times each of the seven desired expressions (Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral), and a series of independent Japanese evaluators gave the images a label, to validate the expressions, however, each of these evaluators was given a set of adjectives to be identified in each image, heavily biasing the categorization of the images. The images are presented centralized and in greyscale, and given the dataset size and the reduced amount

of training images, this dataset will help us to evaluate a very specific scenario. In our experiments, we follow the proposed evaluation scheme of leave-one-emotion-out and calculate the models' accuracy. The decision-making layer for this dataset is also composed of a fully connected layer followed by a SoftMax layer, to provide a one-hot-encoding classification.

The MaskedAffectNet [9] dataset represents a constrained interaction scenario. It is composed of the same images of the AffectNet dataset, but with the artificial addition of a facial mask. The mask is added in a postprocessing scheme that finds the facial points of the mouth. It then uses a geometrical transformation on a standard face mask image and fixes the mask on top of the mouth. The results closely resemble mask-use in a real-world environment.

The OMG-Emotion [5] dataset contains around 10 hours of recordings from persons performing monologues. Each of the 675 videos has a single person and contains about one minute in length. The study collected the videos from the web and they were manually annotated by an internet crowd using an arousal and valence scale. This dataset contains a very specific world representation, as each video has a unique person expressing a continually changing emotional behaviour across a certain topic, so there exists a gradual transition of expressions. The labelling process, although developed from different persons, is based on utterances. This means that a sequence of frames represents the entire labelling scheme, instead of relying on facial expressions alone. Benchmarking facial expression recognition models with this dataset is challenging, as the individual and unique components of how each person expresses emotions are present on the video. The dataset is available in the form of video files, and we pre-process them by cropping faces using the OpenCV face localizer [12]. The authors propose a specific training and validation separation, and we use CCC per arousal and valence as a main performance metric. The decision-making layer is composed of a GRU layer to process sequences, followed by a fully-connected layer and two linear outputs – one for arousal and one for valence.

The ElderReact [56] dataset has 1,323 videos of 46 elderly individuals, all collected from the internet. Each video contains one person naturally expressing emotions. Each video has a few seconds of length and is annotated with the presence or absence of seven affective states: Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, and Frustration. Each video has eight binary labels, one for each affective state. We process each video by cropping the face, using the OpenCV face localizer [12], and training the models using sequential decision-making. Like the OMG-Emotion, we use a GRU layer, followed by a fully connected layer and a SoftMax output layer for each affective state. The dataset authors provide specific training/validation separation that we use in our experiments. We calculate the F1-score between all the affective states as our main performance.

The EmoReact [64] dataset is similar to the ElderReact in construction and labeling proceeding but is composed of

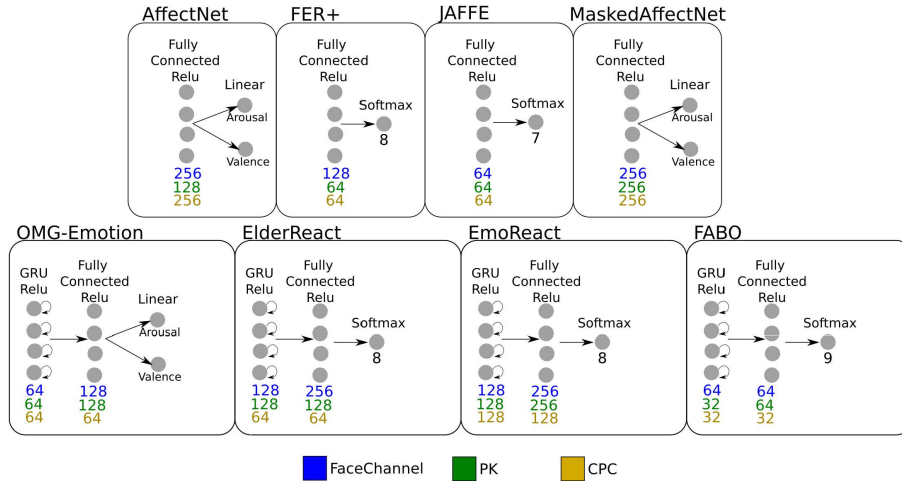


FIGURE 7. Decision making layers of all networks investigated in our experiments. For each evaluated dataset, we implement, optimize and experiment with an individual decision making layer.

videos from children. A total of 1200 videos are available, all of them with a few seconds, and annotated using the same categories present on the ElderReact. Also for this dataset, we use the same sequential decision-making, composed of one GRU, one fully-connected, and one softmax layer per affective category. We use the available training/validation separation in our experiments.

The **FABO** [33] dataset is our last experimental scenario. It contains short videos of actors performing expressions by request. The dataset has a total of 284 videos, each containing 2 to 4 executions of the same expression. Each execution starts from a neutral position followed by the facial expression apex. There is then a return to the neutral position. Each video is labeled using one out of 9 expressions (Anger, Anxiety, Boredom, Disgust, Fear, Happiness, Puzzlement, Sadness, and Surprise) associated with the apex of each video. We process each video by extracting the face using the OpenCV face localizer [12], then feed the apex of each sequence to a sequential decision-making layer with the same structure as the EmoReact and ElderReact models. We use the given training and validation sets and calculate the accuracy as our main performance metric.

B. EVALUATION METRICS

The AffectNet, MaskedAffectNet, FER+, EmoReact, ElderReact, and OMG-Emotion datasets have a standard separation between training and validation samples, which we follow in all our experiments. The JAFFE evaluation follows a leave-one-emotion-out classification scheme, which is the most common evaluation metric in the literature. The FABO dataset follows this as well.

The AffectNet, MaskedAffectNet, and the OMG-Emotion datasets are evaluated in terms of concordance correlation coefficient (CCC) [48] for both arousal and valence representations. The CCC is computed as:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4}$$

where μ_x and μ_y represent the mean for model predictions and the annotations and σ_x^2 and σ_y^2 , are the corresponding variances. ρ is Pearson’s Correlation Coefficient between model prediction labels and the annotations.

The FER+, JAFFE, and FABO datasets use accuracy as the main performance metric, while the EmoReact and ElderReact datasets use the F1-Score averaged per emotional category.

We ran each of our experiments 30 times, and we calculated the average performance, exhibiting it herein. We pre-trained each of the models with the AffectNet dataset, and the facial expression encoders are then used as input to the decision-making layers. The final performance of each model is calculated using a combination of the facial encoder and decision-making.

V. RESULTS

A. AffectNet BASELINE

Our first experimental setting calculates the performance of each model when fully trained with the AffectNet dataset. Figure 8 reports the final performance. The FaceChannel shows slightly better performance on valence, reaching a CCC of 0.46, while the CPC encoder achieves the best arousal with 0.63. In general terms, the performance of all three encoders was similar, showing that all three encoders do learn efficient facial expression representations.

B. FACIAL REPRESENTATION PERFORMANCE

The entire experimental result for the MaskedAffectNet and OMG-Emotion, in terms of CCC, are reported in Figure 9. In general terms, the best results on all four training settings were achieved by fine-tuning all the layers of the network, which is somehow expected, as both datasets have a large amount of data. In our recent published paper [9], we report a similar experiment with the MaskedAffectNet and the FaceChannel encoder and obtained the same results.

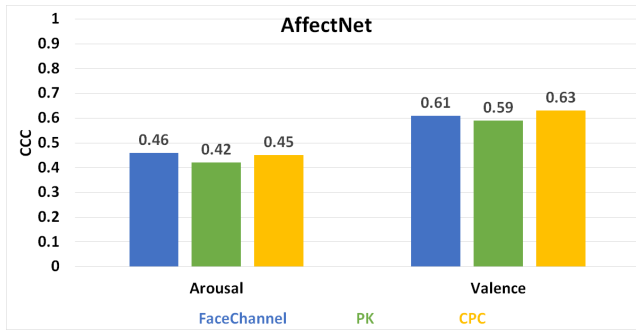


FIGURE 8. Baseline experiment where each of the models (the FaceChannel, the PK network and the CPC network) were trained with the AffectNet dataset. The performance is reported in terms of Concordance Correlation Coefficient (CCC) for arousal and valence.

Training only the decision-making layer presented the worst performance on the MaskedAffectNet dataset, which could indicate that the emotional representation learned from the AffectNet dataset was not enough. This is somehow expected, given the presence of the masks covering much of the faces in this dataset. For the OMG-Emotion, training from the scratch presented the worst results, which leads to the understanding that the dataset alone does not have enough data samples to train these encoders. In both cases, when the Face-STN is present, the results improve drastically, in some cases surpassing the total fine-tuning routine. Both datasets have their own labeling process, and thus, affective representation. The high performance achieved by the Face-STN is clear indication that it can focus the general features learned by the encoders into very specific affective representation.

Similar behavior can be found when evaluating the accuracy-based datasets (FABO, FER+, and JAFFE), reported in Figure 9. The JAFFE dataset is quite particular here because the PK encoder seems to not be able to perform as well as the other two encoders. Probably an indication that the facial representations depicted by the encoder are not enough for the very specific characteristics from the JAFFE dataset. Again, the presence of the Face-STN improves drastically the performance of all encoders. In evaluating the models on the ElderReact and EmoReact datasets that we report in Figure 11, we observe that full retraining obtains the best results, while exclusively training decision-making achieves the worst results. In terms of encoder, all three models achieve similar results.

The presence of the Face-STN also impacts positively the encoders when evaluated with the ElderReact and EmoReact datasets, reported in Figure 11. These datasets show the least variance in the performance range between all the experiments, which shows that their facial representation is not heavily affected by the facial features coming from the encoders.

C. STATE-OF-THE-ART COMPARISON

The Face-STN achieves a competitive performance compared to the current state-of-the-art results on the OMG-Emotion dataset [23], [66], [81], as Table 1 exhibits. All the reported

models use deep neural networks with strong pretraining and fine-tuning routines. Using attention mechanisms [81] to process the continuous expressions in the videos presented the best results of the challenge, such as achieving a CCC of 0:35 for arousal and 0:49 for valence. Temporal pooling, implemented as bi-directional LSTM, achieved the second best, with a CCC of 0:24 for arousal and 0:43 for valence. Late-fusion of facial expressions, speech signals, and text information reached the third-best result, with a CCC of 0:27 for arousal and 0:35 for valence. The complex attention-based network proposed by Huang *et al.* [37] achieved a CCC of 0:31 in arousal and 0:45 in valence, using only visual information. Our Face-STN achieve a maximum of 0.38 arousal (with the PK encoder), and 0.44 valence (with the CPC encoder) without needing to retrain the convolutional layers, reducing the fine-tuning effort.

TABLE 1. CCC, for arousal and valence when evaluating the Face-STN together with the FaceChannel (FC), PK and CPC encoders compared with state-of-the-art models on the OMG-Emotion dataset.

Model	Arousal	Valence
Zheng <i>et al.</i> [81]	0.35	0.49
Huang <i>et al.</i> [37]	0.31	0.45
Peng <i>et al.</i> [66]	0.24	0.43
Deng <i>et al.</i> [23]	0.27	0.35
Face-STN (FC)	0.32	0.43
Face-STN (PK)	0.38	0.41
Face-STN (CPC)	0.36	0.44

Table 2 presents the results of training and evaluating with the FER+ model, wherein our Face-STN achieved better results compared to the dataset authors [10]. They focus on using a fine-tuned VGG13 encoder that updates all the convolutional layers. We also outperform the results Miao *et al.* [59], Li *et al.* [50], and Siqueira *et al.* [71] reported, all of which employ different types of complex neural networks to learn facial expressions. On the FABO dataset, the Face-STN achieves higher results than reported in the literature, including Chen *et al.* [17], who proposed a frame-based recognition and a bag-of-words-based model, or even Gunes *et al.* [32] who used an SVM-based implementation.

When evaluated on the JAFFE dataset, the Face-STN attached to the FaceChannel achieves the best results when compared with the fine-tuning of the DeepEmotion [60] and the attention-based salient patch neural network [34].

The performance of the Face-STN on the EmoReact and ElderReact, reported in Table 3, is better than the models reported by the authors. On the EmoReact, the FaceChannel encoder achieves the best results, while on the ElderReact, the CPC encoder has the highest F1-Score.

Some of the models with which these datasets were evaluated seem to be outdated for other computer vision tasks. In our experiments, however, we do evaluate the performance of three very recent deep neural networks (FaceChannel, PK, and CPC) on each of the datasets. The Face-STN complements these models and presents results that are competitive with them.

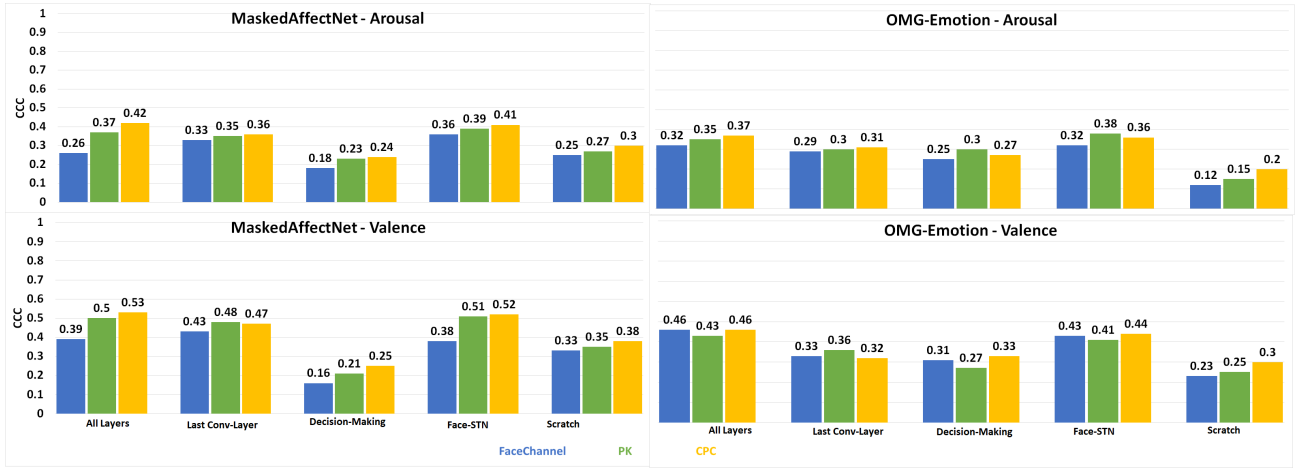


FIGURE 9. Performance in terms of Concordance Correlation Coefficient (CCC), when evaluating the three encoders (FaceChannel, PK, and CPC) on the MaskedAffectNet and OMG-Emotion datasets in five settings: Training the entire encoder and the decision-making layer (All layers), training the last-convolutional layer of the encoder and the decision-making layer (Last Conv-Layer), training only the decision-making layer (Decision-Making), training the Face-STN and the decision making layer (STN-Face), and training the entire network from the scratch (Scratch).

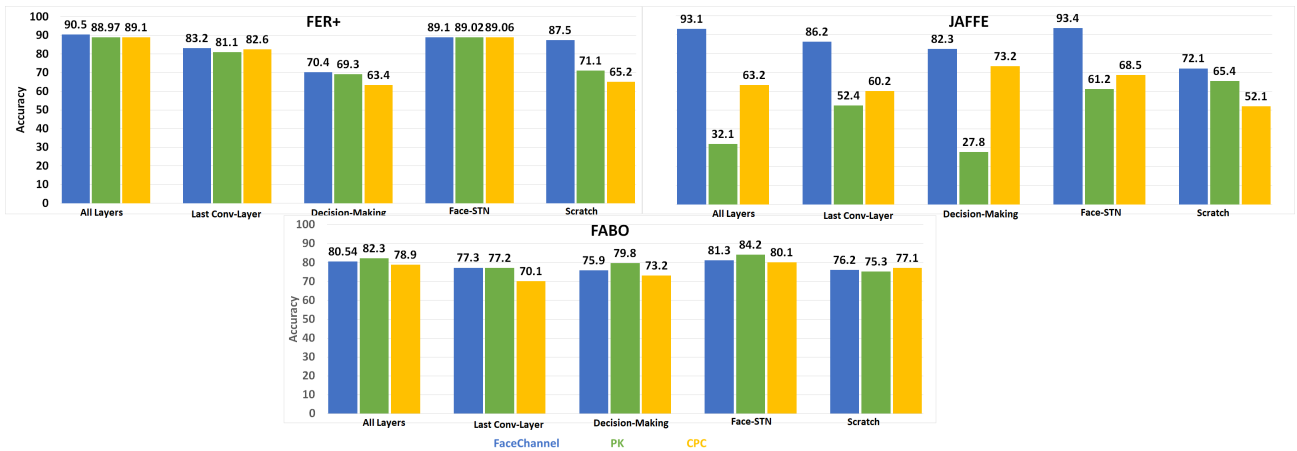


FIGURE 10. Performance, in terms of accuracy, when evaluating the three encoders (FaceChannel, PK and CPC) on the FER+, JAFFE and FABO datasets in five settings: Training the entire encoder and the decision-making layer (All layers), training the last-convolutional layer of the encoder and the decision-making layer (Last Conv-Layer), training only the decision-making layer (Decision-Making), training the Face-STN and the decision making layer (STN-Face), and training the entire network from the scratch (Scratch).

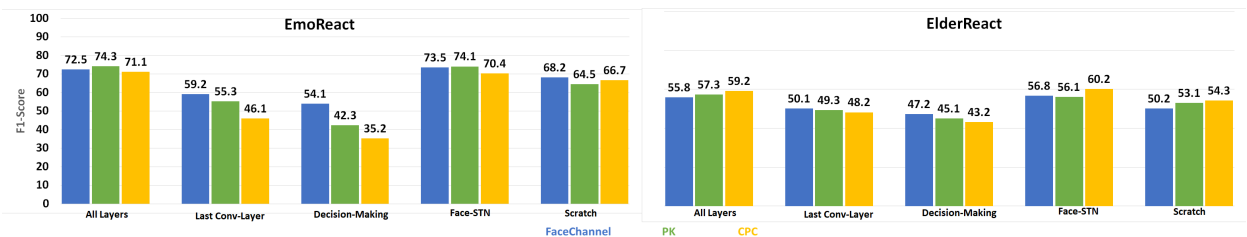


FIGURE 11. Performance, in terms of F1-Score, when evaluating the three encoders (FaceChannel, PK and CPC) on the EmoReact and ElderReact datasets in five settings: Training the entire encoder and the decision-making layer (All layers), training the last-convolutional layer of the encoder and the decision-making layer (Last Conv-Layer), training only the decision-making layer (Decision-Making), training the Face-STN and the decision making layer (STN-Face), and training the entire network from the scratch (Scratch).

VI. DISCUSSIONS

Our experimental results confirm that the Face-STN can be used to adapt different facial encoders towards specific affective worlds. To obtain a holistic understanding of the impact

of the plugins, we must disentangle their training efforts from their final performance and contrast this information with all other training settings. Besides performance, understanding the impact of the Face-STN on the facial representation

TABLE 2. Accuracy when evaluating the Face-STN together with the FaceChannel (FC), PK and CPC encoders compared with state-of-the-art models on the FER+, FABO and JAFFE datasets.

FER+		FABO	
Model	Accuracy	Model	Accuracy
CNN VGG13 [10]	84.98 %	Temp. Norm. [17]	66.5 %
SHCNN [59]	86.54 %	Bag of Words [17]	59.00 %
TFE-JL [50]	84.30 %	SVM [32]	32.49 %
ESR-9 [71]	87.15 %	Adaboost [32]	35.22 %
Face-STN (FC)	89.10 %	Face-STN (FC)	81.30 %
Face-STN (PK)	89.02 %	Face-STN (PK)	84.20 %
Face-STN (CPC)	89.06 %	Face-STN (CPC)	80.10 %
JAFFE			
Model	Accuracy		
DeepEm [60]	92.2 %		
SalientPatch [34]	91.8 %		
Face-STN (FC)	93.40 %		
Face-STN (PK)	61.20 %		
Face-STN (CPC)	68.50 %		

TABLE 3. F1-Score, when evaluating the Face-STN together with the FaceChannel (FC), PK and CPC encoders compared with state-of-the-art models the EmoReact and ElderReact datasets.

EmoReact		ElderReact	
Model	F1-Score	Model	F1-Score
RBFSVM [64]	69.2	SVM [57]	45.8
SVM [64]	66.1	XGBoost [57]	42.6
Face-STN (FC)	75.5	Face-STN (FC)	56.8
Face-STN (PK)	74.1	Face-STN (PK)	56.1
Face-STN (CPC)	70.4	Face-STN (CPC)	60.2

of each encoder is needed to ground its true contributions. We perform feature formation analyses, especially on the representation of very specific affective worlds.

Our main contribution of this paper regards the connection between the Face-STN and the non-universal perception of affect theory. Our experimental setup initially indicates how we can continue this quest, and we further discuss current advantages and limitations of this approach.

A. TRAINING EFFORT VS PERFORMANCE

When we analyse performance alone, our experiments show that retraining all the encoders, in a full training setting, increases drastically the performance in all the datasets. Although it is relatively easy to obtain computational power on demand to train large and complex models, the number of trainable parameters of a model continues to indicate the training effort this model takes to be updated. When comparing the relative performance and number of parameters from the full-training setting and the Face-STNs for each encoder type in each dataset, displayed in Table 4, we observe that the Face-STNs outperform the full training in most datasets.

For most cases, we observe that the Face-STN has a similar performance when compared to the full-training, but even in the worst case, the relative performance achieved by the Face-STN is over 93% of the full-training performance. On the other hand, the training effort, represented by the number of updatable parameters, drastically reduces, especially for the FaceChannel. In the extreme case of the

PK encoder with the JAFFE dataset, the Face-STN achieved almost double the performance.

Besides discussing the numbers and performance bits, the Face-STN displayed an important behaviour that is lacking in most automatic affective perception models: fast adaptability. It could, based on prior perception models (the pretrained encoders), modify the affective representations, embedded on the latent space of each encoder, towards the specific characteristics of each dataset. By doing so, we could reuse the encoder in every dataset without retraining or readapting them. When we did the same by retraining only the convolutional layer, the performance dropped considerably, and we were modifying these encoders drastically, needing one set of encoders per dataset. When not readapting the encoders at all, only updating the decision-making layer, the performance dropped to the lowest levels, making this option the worst of our experiments.

B. HOW THE FACE-STN ALLOWS AFFECTIVE BIASING?

Our results show that the Face-STN networks are able to improve performance in most cases, or at least match the performance of full-retraining on each of the datasets. The main contribution of the Face-STN involves using a bottom-up training scheme to try to adapt the last convolutional layers towards the unique affective characteristics that each of the datasets possesses. That means the affective information coming from the labelling scheme of each dataset directly impacts the selection of specific features that each encoder can extract.

The Face-STN does not update the weights of the last convolutional layer, but it rearranges the features to highlight the most important ones for that specific dataset. If the original encoders, trained on the AffectNet dataset, already have a similar representation to the ones found on the images from a dataset, the impact of the Face-STN is reduced, as one can see in the case of the FABO dataset. When the facial representations learned by the encoders differ from the ones present on the dataset, which is the case of the JAFFE images, the Face-STN could repurpose the learned representations to fit the JAFFE requirements. To illustrate this behaviour better, Figure 12 displays the differences between the entangled representations of the JAFFE dataset of the three encoders when training all layers and Face-STN settings. The entangled representations are passed through a t-SNE calculation to obtain the two most important components. For the PK and CPC encoders, training all layers does not produce distinguishable representations, while when the Face-STN is present, the representations are rearranged and better distinguishable from each other, based on their original labels.

C. AND WHY ARE FACE-STNs NON-UNIVERSAL?

Independently of the affective representations with which we are dealing, faces do not change. The general physical structure and characteristics of a face endure, which is a good start for artificial facial expression recognition because they can focus on which features to adapt. Convolutional

TABLE 4. Relative performance and number of parameters when training and evaluating the Face-STNs, for all encoders, compared to retraining the full network, which usually achieved the best results, for all datasets.

Datasets	Perform. Face-STN \ Full Training			Number Param. Face-STN \ Full Training		
	FaceChannel	PK	CPC	FaceChannel	PK	CPC
MaskedAffectNet Arousal	1.38	1.05	0.97			
MaskedAffectNet Valence	0.97	1.02	0.98			
OMG-Emotion Valence	1.00	1.08	0.97			
OMG-Emotion Arousal	0.93	0.95	0.95			
FER+	0.98	1.00	99.90	0.006	0.022	0.336
JAFFE	1.00	1.90	1.08			
FABO	1.00	1.02	1.01			
EmoReact	1.01	0.99	0.99			
ElderReact	1.01	0.97	1.01			

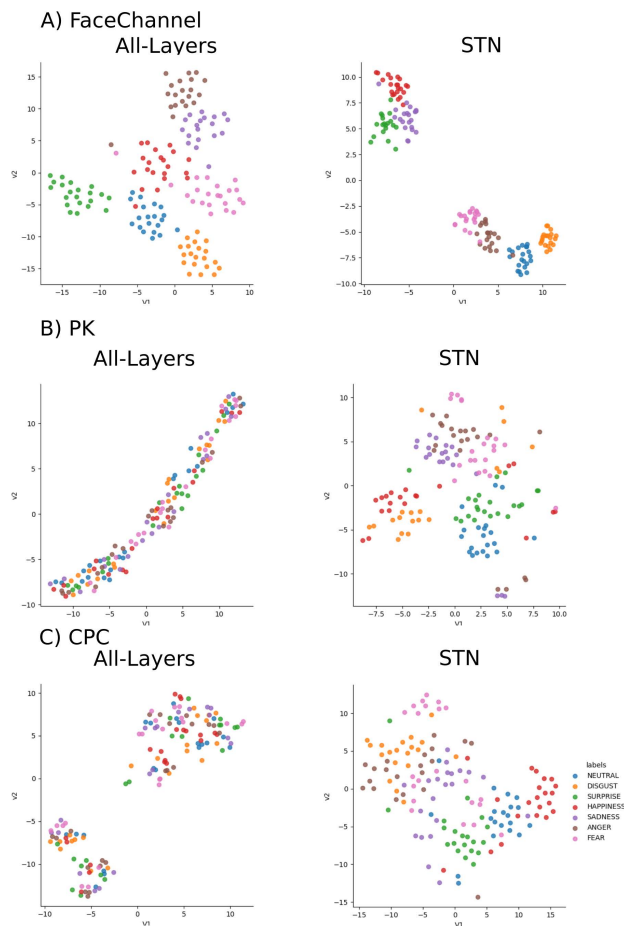


FIGURE 12. Visualization of the entangled representation of all images of the JAFFE dataset when represented by the FaceChannel (A), PK (B), and CPC (C) using the training all layers (ALL) setting and the Face-STN (STN) setting.

neural networks can depict facial characteristics quite well [44], [74], but because they learn it using a strongly supervised process, the given labels still bias the learned representation. This was the case in our experiments in terms of a performance drop, especially when evaluating the pre-trained encoders on very specific affective world representations, such as the JAFFE, the EmoReact, and the ElderReact datasets.

Fitting our experiments within the concept of non-universality of emotional perception can seem contradictory,

as our model focuses on rearranging pretrained perception towards very well-defined-by strong labels-affective worlds. However, the adaption that the Face-STNs achieve allows a pre-existing perception model to deal with unknown conditions from different datasets. Our experiments demonstrate that our model addresses the problem of learning facial representations by reorganizing the existing facial features. It does by biasing the high-level representations towards the labels of each dataset, improving the overall model’s performance. This demonstrates that, at least for a well-defined encoder, different scenarios can share the learned features. By biasing these features we demonstrate that in most cases it is a more beneficial solution than retraining the encoders even partially. Understanding this problem as a continual rearranging of a perception mechanism, based on the specific affective context given by each dataset, is how we address the non-universality of emotional perception.

VII. CONCLUSION AND FUTURE WORK

In this paper, we present a facial expression perception study where we investigate the readaptation of facial features as a mechanism for achieving non-universal affective perception. In this regard, we present a Spatial Transformer Network (Face-STN) that one may attach to any convolution-based encoder to rearrange learned features without the need of retraining the entire encoder. We perform a series of experiments with three different convolution-based encoders and with eight different datasets, representing different affective worlds. Our experiments demonstrate that when the Face-STNs are present, we reduce the training effort and maintain high performance, sometimes even surpassing the state-of-the-art performance on each of the evaluated datasets.

Besides performance, we discuss how our Face-STNs adapt the concept of non-universal emotional perception and put it into practice by understanding its impact on the different affective representations of each dataset. We establish and present our networks as one tool that will help us approach non-universal perception in affective computing, which will help develop truly adaptable emotional perception models.

Furthermore, the major contribution of this study regards discussing the impact and responsibility when developing facial expression research. We should consider the soft

separation of face representation from affect understanding, following the recent trend on affective perception of humans, to provide reliable and adaptable facial expression recognition solutions. Focusing on adaptable affective recognition, instead of a general one, will allow us to be much more flexible when dealing with underrepresented scenarios.

Although we demonstrate the capability of the Face-STNs to adapt towards very specific affective worlds, we are still dealing with perception alone. All our experiments consider as granted that the labels derived from the datasets are reliable and represent the truth of that affective scenario. In future work, we will continue our search for non-universal emotional modelling from the affective understanding perspective, primarily addressing the problems of emotional grounding in different scenarios. We will address this problem by adapting the Face-STN to consider other aspects of the scenario, such as using reinforcement learning to address the congruence of the affective responses of a person.

**APPENDIX A
DECISION-MAKING NETWORKS**

For each of the datasets, we propose one decision-making network that is attached to each of the encoders. The final architecture, and topological and training parameters, of these networks were found using a tree-parzen search [11] through the search space found in Table 5.

TABLE 5. Search space used to optimize all of our decision-making networks for all datasets.

Parameter	Search Space
AffectNet, FER+, JAFFE, MaskedAffectNet	
# Dense Layers	[1,2,3]
# Neurons Per Layer	[16, 32, 64, 128, 256, 512, 1024]
Optimizer	[SGD, ADAM]
Learning Rate	[0.0005; 0.9]
CIAO temperature	[0.001; 0.09]
OMG-Emotion, EmoReact, ElderReact, FABO	
# Sequence Length	[5,10,15,30]
# Dense Layers	[1,2,3]
# GRU Layers	[1,2,3]
# Neurons Per Layer	[16, 32, 64, 128, 256, 512, 1024]
Optimizer	[SGD, ADAM]
Learning Rate	[0.0005; 0.9]
CIAO temperature	[0.001; 0.09]

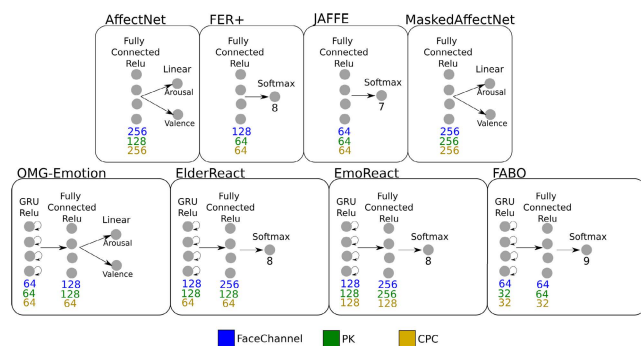


FIGURE 13. Final architecture of each decision-making layer for each of the datasets.

TABLE 6. Final architecture of each decision-making network for each of the datasets.

Parameter	FaceChannel	PK	CPC
AffectNet			
Dense Layers	1	1	
Neurons Per Layer	256	128	256
Optimizer	SGD	SGD	SGD
Learning Rate	0.08	0.06	0.08
CIAO Temperature	0.03	0.02	0.03
FER+			
Dense Layers	1	1	1
Neurons Per Layer	128	64	64
Optimizer	SGD	SGD	SGD
Learning Rate	0.002	0.006	0.006
CIAO Temperature	0.23	0.23	0.23
JAFFE			
Dense Layers	1	1	1
Neurons Per Layer	64	64	64
Optimizer	SGD	SGD	SGD
Learning Rate	0.001	0.005	0.001
CIAO Temperature	0.4	0.4	0.4
MaskedAffectNet			
Dense Layers	1	1	
Neurons Per Layer	256	256	256
Optimizer	SGD	SGD	SGD
Learning Rate	0.08	0.08	0.08
CIAO Temperature	0.03	0.03	0.03
OMG-Emotion			
# Sequence Length	15	15	15
# Dense Layers	1	1	1
# LSTM Layers	1	1	1
# Neurons Dense Layer	128	128	64
# Neurons GRU Layer	64	64	64
Optimizer	SGD	SGD	SGD
Learning Rate	0.4	0.2	0.2
CIAO Temperature	0.4	0.4	0.4
ElderReact			
# Sequence Length	15	15	15
# Dense Layers	1	1	1
# LSTM Layers	1	1	1
# Neurons Dense Layer	256	128	64
# Neurons LSTM Layer	128	128	64
Optimizer	SGD	SGD	SGD
Learning Rate	0.4	0.2	0.08
CIAO Temperature	0.4	0.4	0.32
EmoReact			
# Sequence Length	15	15	15
# Dense Layers	1	1	1
# LSTM Layers	1	1	1
# Neurons Dense Layer	256	256	128
# Neurons LSTM Layer	128	128	128
Optimizer	SGD	SGD	SGD
Learning Rate	0.1	0.1	0.2
CIAO Temperature	0.2	0.2	0.15
FABO			
# Sequence Length	9	9	9
# Dense Layers	1	1	1
# LSTM Layers	1	1	1
# Neurons Dense Layer	64	64	32
# Neurons LSTM Layer	64	32	32
Optimizer	SGD	SGD	SGD
Learning Rate	0.05	0.07	0.05
CIAO Temperature	0.2	0.2	0.2

The final architecture of each decision-making network is reported in Table 6, and illustrated in Figure 13.

REFERENCES

[1] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharruddin, "Survey on deep neural networks in speech and vision systems," *Neurocomputing*, vol. 417, pp. 302–321, Dec. 2020.

- [2] S. F. Aly and A. L. Abbott, "Facial emotion recognition with varying poses and/or partial occlusion using multi-stage progressive transfer learning," in *Proc. Scand. Conf. Image Anal.* Cham, Switzerland: Springer, 2019, pp. 101–112.
- [3] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," 2020, *arXiv:2007.16104*.
- [4] P. Barros, E. Barakova, and S. Wermter, "Adapting the interplay between personalized and generalized affect recognition based on an unsupervised neural framework," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1349–1365, Jul. 2022.
- [5] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-emotion behavior dataset," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.
- [6] P. Barros, N. Churamani, and A. Sciutti, "The FaceChannel: A fast and furious deep neural network for facial expression recognition," *Social Netw. Comput. Sci.*, vol. 1, no. 6, pp. 1–10, Nov. 2020.
- [7] P. Barros, G. Parisi, and S. Wermter, "A personalized affective memory model for improving emotion recognition," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 485–494.
- [8] P. Barros, G. I. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: A deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, Aug. 2017.
- [9] P. Barros and A. Sciutti, "I only have eyes for you: The impact of masks on convolutional-based facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1226–1231.
- [10] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.
- [11] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: A Python library for model selection and hyperparameter optimization," *Comput. Sci. Discovery*, vol. 8, no. 1, 2015, Art. no. 014008.
- [12] G. Bradski and A. Kaehler, "OpenCV," *Dr. Dobbs's J. Softw. Tools*, vol. 3, p. 120, Nov. 2000.
- [13] I. Buciu, C. Kotropoulos, and I. Pitas, "ICA and Gabor representation for facial expression recognition," in *Proc. Int. Conf. Image Process.*, 2003, p. 855.
- [14] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep convolutional neural network for expression recognition," 2015, *arXiv:1509.05371*.
- [15] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, S. Han, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," 2019, *arXiv:1903.08051*.
- [16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
- [17] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," *Image Vis. Comput.*, vol. 31, no. 2, pp. 175–185, Feb. 2013.
- [18] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. Chen, "Linear regression-based adaptation of music emotion recognition models for personalization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2149–2153.
- [19] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen, "Component tying for mixture model adaptation in personalization of music emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 7, pp. 1409–1420, Jul. 2017.
- [20] N. Churamani and H. Gunes, "CLIFER: Continual learning with imagination for facial expression recognition," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2020, pp. 322–328.
- [21] J. F. Cohn, Z. Ambadar, and P. Ekman, "Observer-based measurement of facial expression with the facial action coding system," *Handbook Emotion Elicitation Assessment*, vol. 1, no. 3, pp. 203–221, 2007.
- [22] A. S. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, and G. Prasad, "Sixteen facial expressions occur in similar contexts worldwide," *Nature*, vol. 589, no. 7841, pp. 251–257, 2020.
- [23] D. Deng, Y. Zhou, J. Pi, and B. E. Shi, "Multimodal utterance-level affect analysis using visual, audio and text features," 2018, *arXiv:1805.00625*.
- [24] S. Dubuisson, F. Davoine, and M. Masson, "A solution for facial expression representation and recognition," *Signal Process., Image Commun.*, vol. 17, no. 9, pp. 657–673, Oct. 2002.
- [25] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.
- [26] P. Ekman and H. Oster, "Facial expressions of emotion," *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, 1979.
- [27] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford Univ. Press, 1997.
- [28] M. J. Farah, K. D. Wilson, M. Drain, and J. N. Tanaka, "What is 'special' about face perception?" *Psychol. Rev.*, vol. 105, no. 3, p. 482, 1998.
- [29] L. F. Barrett, "AI weighs in on debate about universal facial expressions," *Nature*, vol. 589, no. 7841, pp. 202–203, Jan. 2021.
- [30] Y. Frégnac, C. Monier, F. Chavane, P. Baudot, and L. Graham, "Shunting inhibition, a silent step in visual cortical computation," *J. Physiol.-Paris*, vol. 97, nos. 4–6, pp. 441–451, Jul. 2003.
- [31] J. Gao, Y. Fu, Y.-G. Jiang, and X. Xue, "Frame-transformer emotion classification network," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 78–83.
- [32] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 64–84, Feb. 2009.
- [33] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 1148–1153.
- [34] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Mar. 2015.
- [35] K. Hoemann, C. Nielson, A. Yuen, J. W. Gurera, K. S. Quigley, and L. F. Barrett, "Expertise in emotion: A scoping review and unifying framework for individual differences in the mental representation of emotional experience," *Psychol. Bull.*, vol. 147, no. 11, p. 1159, 2021.
- [36] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using Wasserstein distance for speech enhancement," 2020, *arXiv:2010.15174*.
- [37] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5866–5870.
- [38] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, Sep. 2019.
- [39] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara, "Cultural confusions show that facial expressions are not universal," *Current Biol.*, vol. 19, no. 18, pp. 1543–1548, Sep. 2009.
- [40] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 366, no. 6472, pp. 1517–1522, Dec. 2019.
- [41] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.
- [42] N. Kanwisher, "Domain specificity in face perception," *Nature Neurosci.*, vol. 3, no. 8, pp. 759–763, Aug. 2000.
- [43] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 19–27.
- [44] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 223–236, Apr. 2019.
- [45] A. Kirchner, M. Boiger, Y. Uchida, V. Norasakkunkit, P. Verduyn, and B. Mesquita, "Humiliated fury is not universal: The co-occurrence of anger and shame in the United States and Japan," *Cognition Emotion*, vol. 32, no. 6, pp. 1317–1328, Aug. 2018.
- [46] J. Kumari, R. Rajesh, and K. M. Pooja, "Facial expression recognition: A survey," *Proc. Comput. Sci.*, vol. 58, pp. 486–491, Jan. 2015.
- [47] M. La Mura and P. Lamberti, "Human-machine interaction personalization: A review on gender and emotion recognition through speech analysis," in *Proc. IEEE Int. Workshop Metrology Ind. IoT*, Jun. 2020, pp. 319–323.
- [48] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
- [49] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 391–404, Feb. 2020.

- [50] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, and X. Li, "Facial expression recognition with identity and emotion joint learning," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 544–550, Apr. 2021.
- [51] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022.
- [52] S.-T. Liong, Y. Gan, D. Zheng, S.-M. Li, H.-X. Xu, H.-Z. Zhang, R.-K. Lyu, and K.-H. Liu, "Evaluation of the spatio-temporal features and Gan for micro-expression recognition system," *J. Signal Process. Syst.*, vol. 92, no. 7, pp. 705–725, 2020.
- [53] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.
- [54] M. J. Lyons, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets (IVC special Issue)," 2020, *arXiv:2009.05938*.
- [55] C. Ma, L. Chen, and J. Yong, "AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection," *Neurocomputing*, vol. 355, pp. 35–47, Aug. 2019.
- [56] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency, "ElderReact: A multimodal dataset for recognizing emotional response in aging adults," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 349–357.
- [57] M. Marini, A. Ansani, F. Paglieri, F. Caruana, and M. Viola, "The impact of facemasks on emotion recognition, trust attribution and re-identification," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, Dec. 2021.
- [58] D. Mehta, M. Siddiqui, and A. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality," *Sensors*, vol. 18, no. 2, p. 416, Feb. 2018.
- [59] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78000–78011, 2019.
- [60] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," 2019, *arXiv:1902.01019*.
- [61] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2017.
- [62] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes, and S. Wermter, "Understanding how deep neural networks learn face expressions," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 227–234.
- [63] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 443–449.
- [64] B. Nojavanasghari, T. Baltrusaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: A multimodal approach and dataset for recognizing emotional responses in children," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 137–144.
- [65] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
- [66] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler, "A deep network for arousal-valence emotion prediction with acoustic-visual cues," 2018, *arXiv:1805.00638*.
- [67] D. Reichardt, "Affective computing needs personalization—And a character?" in *Character Computing*. Cham, Switzerland: Springer, 2020, pp. 87–98.
- [68] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer*, 2012, pp. 808–822.
- [69] R. Savery and G. Weinberg, "A survey of robotics and emotion: Classifications and models of emotional interaction," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 986–993.
- [70] M. Sidorov, S. Ultes, and A. Schmitt, "Emotions are a personal thing: Towards speaker-adaptive emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4803–4807.
- [71] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," 2020, *arXiv:2001.06338*.
- [72] C. Sumathi, T. Santhanam, and M. Mahadevi, "Automatic facial expression analysis a survey," *Int. J. Comput. Sci. Eng. Surv.*, vol. 3, no. 6, p. 47, 2012.
- [73] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2018, pp. 270–279.
- [74] C.-H. Tang, G.-S.-J. Hsu, and M. H. Yap, "Face recognition with disentangled facial representation learning and data augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1670–1674.
- [75] D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception," *Annu. Rev. Neurosci.*, vol. 31, no. 1, pp. 411–437, Jul. 2008.
- [76] 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, May 2017, Art. no. e0177239.
- [77] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, and L. Xie, "A novel feature separation model exchange-GAN for facial expression recognition," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106217.
- [78] A. W. Young, D. Hellawell, and D. C. Hay, "Configurational information in face perception," *Perception*, vol. 42, no. 11, pp. 1166–1178, Nov. 2013.
- [79] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, Jan. 2018.
- [80] T. Zhang, "Facial expression recognition based on deep learning: A survey," in *Proc. Int. Conf. Intell. Interact. Syst. Appl.* Cham, Switzerland: Springer, 2017, pp. 345–352.
- [81] Z. Zheng, C. Cao, X. Chen, and G. Xu, "Multimodal emotion recognition for one-minute-gradual emotion challenge," 2018, *arXiv:1805.01060*.



PABLO BARROS received the B.Sc. degree in information systems from the Universidade Federal de Pernambuco, the M.Sc. degree in computer engineering from the Universidade de Pernambuco, Brazil, and the Ph.D. degree in computer science from Universität Hamburg, Germany, in 2016. He worked at several research projects at the University of Hamburg, in particular the Cross-Modal Learning (CML) International Collaboration Consortium, and at the COgNiTive

Architecture for Collaborative Technologies (CONTACT) Unit, Istituto Italiano di Tecnologia (IIT). Currently, he is a Senior Research Scientist with the Sony Research and Development Center, Brussels, Belgium. His main research interests include deep learning and affective computing applied for emotional perception and representation, affect-based human–robot interaction, and its application on social robots.



ALESSANDRA SCIUTTI (Member, IEEE) received the Ph.D. degree in humanoid technologies from the University of Genova, in 2010. She is a Tenure Track Researcher and the Head of the COgNiTive Architecture for Collaborative Technologies (CONTACT) Unit of the Italian Institute of Technology (IIT). After two research periods in USA and Japan, in 2018 she has been awarded the ERC Starting Grant wHiSPER (www.whisperproject.eu), focused on the investigation of joint perception between humans and robots. She published more than 80 papers in international journals and conferences and participated in the coordination of the CODEFROR European IRSES Project. She is currently an Associate Editor of several journals, among which the *International Journal of Social Robotics*, the *IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS*, and *Cognitive Systems Research*. The scientific aim of her research is to investigate the sensory and motor mechanisms underlying mutual understanding in human–human and human–robot interaction. More info at <https://www.iit.it/people/alessandra-sciutti>.

...