IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2023, accepted 29 June 2023, date of publication 10 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293728

RESEARCH ARTICLE

Improving Facial Expression Recognition Through Data Preparation and Merging

CHRISTIAN MEJIA-ESCOBAR[1], MIGUEL CAZORLA[2], (Senior Member, IEEE), AND ESTER MARTINEZ-MARTIN[2], (Senior Member, IEEE)

[1]FIGEMPA, Central University of Ecuador, Quito 170129, Ecuador
[2]Institute for Computer Research, University of Alicante, 03690 Alicante, Spain

Corresponding author: Christian Mejia-Escobar (cimejia@uce.edu.ec)

This work has been funded by grant CIPROM/2021/017 awarded by the Metodología para la educación consciente de las emociones basada en la inteligencia artificial (MEEBAI) Project (Prometheus Programme for Research Groups on R&D Excellence) from Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital of Generalitat Valenciana (Spain), and partially by the grant awarded by the Central University of Ecuador through budget certification No. 34 of March 25, 2022 for the development of the research project with code DOCT-DI-2020-37.

ABSTRACT Human emotions present a major challenge for artificial intelligence. Automated emotion recognition based on facial expressions is important to robotics, medicine, psychology, education, security, arts, entertainment and more. Deep learning is promising for capturing complex emotional features. However, there is no training dataset that is large and representative of the full diversity of emotional expressions in all populations and contexts. Current facial datasets are incomplete, biased, unbalanced, error-prone and have different properties. Models learn these limitations and become dependent on specific datasets, hindering their ability to generalize to new data or real-world scenarios. Our work addresses these difficulties and provides the following contributions to improve emotion recognition: 1) a methodology for merging disparate in-the-wild datasets that increases the number of images and enriches the diversity of people, gestures, and attributes of resolution, color, background, lighting and image format; 2) a balanced, unbiased, and well-labeled evaluator dataset, built with a gender, age, and ethnicity predictor and the successful Stable Diffusion model. Single- and cross-dataset experimentation show that our method increases the generalization of the FER2013, NHFI and AffectNet datasets by 13.93%, 24.17% and 7.45%, respectively; and 3) we propose the first and largest artificial emotion dataset, which can complement real datasets in tasks related to facial expression.

INDEX TERMS Artificial dataset, deep learning, convolutional neural network, emotion recognition, facial expression recognition, stable diffusion.

I. INTRODUCTION

This is the era of artificial intelligence (AI), motivated by increasing computational power, greater availability of data, research and development of ''smart algorithms'', as well as collaborative and interdisciplinary work to address problems in different fields [1]. Numerous AI systems and applications related to computer vision, speech recognition, natural language processing, autonomous driving, healthcare, banking, commerce and other fields demonstrate impressive progress towards imitating human intelligence. Some of our physical and cognitive abilities have been matched and even surpassed, e.g., by stronger and faster robots, translation systems, strategy games and data analysis [2], [3]. However, there are significant challenges in the emotional domain that AI needs to solve to approach the level of human intelligence as a whole.

A. BACKGROUND

Emotions are essential and still distinctive of people, influence our behavior and play a fundamental role in how we communicate and interact with our family, work environment and the rest of society [4]. The perception of emotions is a difficult task for computers, and sometimes even for humans,

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

due to the complex features of each type of emotion and the subtle distinctions between similar emotions. Emotion recognition (ER) has emerged as an active research topic seeking a more natural human-robot interaction (HRI), given the increasing presence of robots in our daily lives [5], [6], [7]. Although social and affective robotics is the predominant application in related literature, the development of systems capable of recognizing emotions is important in areas such as emotional intelligence, medicine, psychology, sociology, public security, video surveillance, road safety, marketing and sales, education, arts and entertainment [8], [9].

Understanding how emotions are expressed and communicated is the key to recognizing them. It is a complex multimodal mechanism involving verbal, nonverbal and corporal components. However, facial expression is the primary and main way to identify human emotions [10], [11], [12]. There is a well-known saying that the face says it all [13]. Even the gesture would be sufficient to describe the emotion we experience, as we often pay more attention to the face than to the words. There is evidence that the patterns of facial expression associated with emotions are universal for all people, in all cultures and contexts [14]. It is very likely that this is for a biological reason [7], [15]. A facial expression involves muscle movements on the human face and the geometric shape of its elements (e.g., eyes, eyebrows, nose, mouth, lips, cheekbones, chin), which are defined in the facial action coding system (FACS) as *action units* (AUs) [16], [17]. A collection of AUs acting together may suggest or reveal a certain type of emotion [13], [18], [19].

Therefore, emotion can be assumed by recognizing the facial expression. This has motivated facial expression recognition (FER) as an important area of research and development. A popular approach to emotion recognition is based on FER from static facial images and videos [8], [9], [20]. This is a difficult task of identification of facial expression features and subsequent classification of emotion, either in dimensional terms, e.g., valence and arousal (how positive or negative an emotion is and the intensity of the emotion, respectively), or in categorical terms, normally one of the seven basic universal emotions: angry, disgust, fear, happy, neutral, sad and surprise [21], [22], [23].

Humans capture emotional changes through the observation of facial expressions. Similarly, computer vision addresses the problematic of FER [5]. The most commonly used techniques are classical machine learning and deep learning, the latter standing out for its ability to automatically extract features from the broad spectrum of emotional gestures through supervised learning [24], [25]. A deep learning solution is composed of a model and data, both elements are fundamental to achieve good results. However, when data quality is poor, no matter how many architectures and configurations of models are tested, improving recognition accuracy is very difficult.

Our work focuses on datasets to train emotion recognition models for real scenarios. A *dataset* is nothing more than a collection of data. In this case, images of human faces captured in any environment or also extracted from videos. At the dimensional level, there are very few facial datasets labeled [22], whereas for a categorical model, a wider range of available datasets can be found. They are collected mainly in two ways [15], [25]: (1) *in-the-lab*, under controlled conditions with actors exaggerating the gestures of each emotion, which allows a high accuracy rate and the FER problem is considered solved, but of little use for real-world applications, and (2) *in-the-wild*, typically datasets collected from the Internet for the purpose of more generalization, whose images include real people in real-world situations. Although there are several in-the-wild datasets for training emotion recognition models, it is very difficult to obtain the same accuracy of controlled environments in uncontrolled environments.

B. PROBLEM

In-the-wild datasets present problems related to lack of quality and generality. In terms of quality, the automatic collection of images from the Internet and the subjectivity of crowdsourcing are the main factors for the presence of irrelevant images, mislabeling by similar and difficult to differentiate emotions, and an imbalance of classes of emotion. In terms of generality, the difficulty of covering all types of human faces and facial expressions results in a bias of features of the general population. For instance, each dataset may contain more images of people of a certain gender, age or ethnicity, which affects recognizing rare or absent emotions in the training dataset. This bias impacts the state-of-the-art in emotion recognition, and the model with the highest accuracy on a specific dataset may not always be the best at recognizing emotions in the real world. In addition, each dataset has its own technical specifications, such as number of images, color, resolution, background, illumination and file type. As a result, the models are highly dependent on the properties of the training dataset, which limits their adaptability to real-world scenarios where facial expressions need to be recognized from images captured in uncontrolled environments.

A large, high-quality dataset that represents the complex heterogeneity of emotional expressions and the entire population in a complete and balanced way would be the ideal resource for training a good-performing recognition model in real-world applications. However, creating a dataset with these requirements is a huge problem [23]. As an approach, our work aims to provide a large dataset, more diverse and general than the existing ones, by merging several in-the-wild FER datasets. A model trained on this merged dataset would generalize more datasets rather than a specific one. To evaluate the model, it is convenient to use images not seen before. Therefore, the dataset is divided into two parts, one of which we will use as a test dataset. This evaluator dataset can also measure the generalization of other models trained on different datasets, which becomes a benchmark for a more generic state-of-the-art and not specific as at present. To achieve this, the evaluator dataset must not only be com-

bined, but also balanced and unbiased, capable of providing quality performance metrics for practical applications.

C. RESEARCH QUESTIONS

The aim of this paper is to answer the following research questions:
RQ1: Can merging several in-the-wild FER datasets improve the generalization of an emotion recognition model?
RQ2: Can a merged, balanced, unbiased dataset evaluate and compare different emotion recognition models, provide a performance metric closer to a general state-of-the-art, and be useful in selecting the most suitable model for real-world applications?

A positive answer to both questions will not completely solve the problem of the lack of quality and generalizability of FER datasets, but our contribution may be one more step that will help future research towards a long-term solution.

D. RESEARCH DIFFICULTIES OF THIS AREA

Computer-based emotion recognition in real-world situations is a major challenge. This has motivated a field of research that must deal with difficulties such as:

- Dependence on advances in other areas. Recognizing emotions is difficult even for people, so there is ongoing work and research by fields such as medicine, psychology and sociology. While humans do not fully understand this topic, endowing machines with this capability will be a problem to solve.
- The emotions we experience individually, in groups and with the rest of society respond to stimuli perceived by our senses that AI lacks and attempts to mimic. A popular approach is DL-based computer vision with images of our face, which is the main indicator, but an emotion is composed of other factors that are difficult to capture in a multimodal training dataset.
- How to measure and catalog human emotions. Dimensional models attempt to approximate more closely the wide range of emotional states of a person; however, they are complex and scarce in terms of available resources. Researchers prefer categorical models because of their simplicity and greater availability; however, the limited number of categories leads to intraclass variability and interclass overlap. Designing an emotion model that takes into account the best of both could lead to better results.

We focused on issues related to the raw material of DL, which is data, both for the training and evaluation phases. Currently, the available training datasets are replete with problems. None is sufficiently large and representative to allow a model to learn patterns that can be generalized to new data and practical situations. The creation of a large dataset of facial images involves their collection and preparation. Whether manual or automated, this process is labor-intensive and time-consuming for large image datasets, as well as error-prone. As datasets grow, so do their problems, becoming difficult to ensure their quality. In emotion recognition, evaluation is also a problem. The ideal assumption that the training and test images come from the same dataset is insufficient. When the training and test samples belong to different datasets, the performance of the FER methods can decrease drastically due to the mismatch in the distribution of features between the two sets. As data-centric research is scarce, contributions in the above two aspects become critical. Our purpose is to design and apply automated strategies to train DL models with better quality data, i.e., AI helping AI.

E. PROPOSAL

Data collection is a costly activity in time and effort, so we combine known FER datasets in a way that acts as a larger and more diverse dataset. The proposed dataset integrates facial images from the FER2013, NHFI and AffectNet datasets. Although combining the datasets expands the range of facial features, it also incorporates the common problems of in-the-wild datasets that limit the improvement of FER models. To deal with the misclassification, we use the reclassified versions of these datasets, obtained using the *data-centric* method presented in our previous work [26]. This method seeks to iteratively refine a dataset, generating a new reclassified version to increase the accuracy of a model. Merging these reclassified datasets results in a more diverse and better classified dataset, the largest merged dataset of FER to our knowledge. A model trained on this dataset becomes the first and most generic one until now. To evaluate it, we design a smaller dataset that meets the requirements of being combined, balanced and unbiased. This evaluator dataset is a more realistic benchmark of the generalization of emotion recognition models and thus of the datasets on which they are trained. We address the problem of imbalance, both among classes and among the relevant characteristics of the population. Facial images are selected according to different categories of gender, age and ethnicity of the persons portrayed [27]. We use a predictor model of these characteristics to assign the images to the corresponding categories. This allows us to choose the same number of images for each category and dataset. Incomplete categories are balanced with synthetic images generated using *Stable Diffusion*, a recently successful AI generative model [28], [29]. The high quality of the synthetic images indicates that this approach is a convenient and effective alternative to traditional techniques such as data augmentation and GAN [30], [31]. This motivates us to create the first and largest fully artificial FER dataset in the world. To answer the questions of this research, we perform a series of experiments following the *single-* and *cross-dataset* approaches, using the real and artificial datasets to train and evaluate customized and pretrained CNNs. This allows us to compare the performance of the models in different contexts.

F. CONTRIBUTION

As DL models learn from data, improving emotion recognition requires improving the training dataset. For this reason,

our core contribution is a novel dataset to train a more robust model for real contexts, instead of the current models based on a specific dataset. To this end, we combine all facial images from the FER2013, NHFI and AffectNet datasets. We use the reclassified versions of these datasets, which are more reliable and allow higher accuracy. To our knowledge, this is the largest merged dataset of categorical emotion covering different image properties.

Additional contributions derived from this main product are:

- A smaller but balanced and unbiased dataset in terms of gender, age and ethnicity for benchmarking the generalization of a FER model and the dataset on which it is trained. This metric is an approximation to a more general state-of-the-art in emotion recognition.
- A novel dataset of synthetic facial images using the state-of-the-art generative tool Stable Diffusion. To our knowledge, this is the first artificial FER dataset, which may lead to a new category in addition to in-the-lab and in-the-wild datasets. We describe prompts engineering designed to control image content and automatic labeling using facial expression action units and the categories of age, gender and ethnicity. This is a useful product for research and development in FER, providing high quality synthetic images in less time compared to traditional techniques, without the need for training and the potential risks of invasion of people's privacy as with real datasets [9], [30].
- A Web-based system for emotion recognition using the model trained on the merged FER dataset. Appendix A presents a use case related to the educational environment. The resources used, the code developed and the products obtained are publicly available at the GitHub.[1]

The content of the paper is structured as follows: Section II reviews related work. Section III describes in detail the materials and methods used to create the merged, evaluator and artificial FER datasets, as well as the CNN-based models. Section IV explains the experimental part, the datasets, the training and the evaluation of the results. Finally, Section V states the conclusions and possible guidelines for future work in Section VI.

II. RELATED WORK

In this section, we present an overview of existing work for combining FER datasets and creating artificial facial datasets, the two central topics of our research. Finally, we highlight the differences and novelties of our work.

A. COMBINATION OF FER DATASETS

The traditional approach to FER is *single-dataset*, which consists of training and evaluating on the same dataset, reserving one part for training data and the other for evaluation data. This limits the generalization capability of a model and its performance in real environments. Our focus is on the *cross-dataset* approach, which attempts to incorporate more datasets into training, evaluation, or both. Few works use and investigate this approach in depth. Ramis et al. [27] consider 4 known datasets (BU-4DFE, CK+, JAFFE and WSEFEP) and 2 new ones (FEGA and FE-test), and discard FER+ and AffectNet due to the problems of in-the-wild datasets. Accuracy reaches 70% when the CNN-based model is trained on the combination of the first five datasets and evaluated in FE-test. As the datasets are collected under different acquisition conditions and diversity of ethnicity and age, each adds important information in training and improves the results over a single dataset approach. Meng et al. [25] use ExpW and Sfew in-the-field, FERPlus and Raf-db in-the-wild, CK+ and Jaffe in-the-lab datasets. These datasets are successively combined from 2 to 6 for training (multiple source) and only one for testing (single target). The cross-dataset result shows a more general CNN-based model, but not necessarily the larger amount of data may improve the accuracy. The combination is because the generalization ability of small datasets should be low. Chaudhari et al. [11] combine FER2013, CK+ and AffectNet into one called AVFER, which is used to train and evaluate Vision Transformer (ViT) and ResNet-18. Accuracy reaches 50.05% and 53.10% for ResNet-18 and ViT, respectively. As one of the first attempts of a transformer for FER, the result is acceptable. Ghosh et al. [9] use FER2013 and CK+ to develop a shared Federated Learning (FL) model. Two client devices and a central server are considered, all with the MobileNet network. The images remain outside the server to ensure privacy. One client trains on FER2013 and the other on CK+. The weights are sent to the server to calculate the averages, which are assigned to the global model. Single- and cross-dataset evaluation shows better MobileNet-fed accuracy (0.7657) and recall (0.7450) indicators, compared to the models trained only on FER2013 or CK+. This is a more robust and safer way to train with facial images. Abou et al. [32] create the 3RL dataset containing 24K facial images from combining FER2013, CK+ and a dataset generated by students.[2] The experiments applying SVM and CNN on the single and combined datasets indicate better generalization using the deep learning method. Kim et al. [33] combine FER2013, CK+ and iSPL.[3] The facial images are standardized using the FIT machine, a multitask cascade neural network and a resizing program. After merging the datasets, the performance of the Xception-based model increases the validation accuracy by 15.33%.

The cited works agree on combining datasets to improve the generalization of a FER model. For the combination, in-the-lab datasets are mostly used, given their high reliability and generally smaller size, whereas FER2013 and AffectNet datasets are rarely considered. In contrast, our work proposes to merge exclusively in-the-wild datasets, previously refined. On the other hand, model training is performed on the combined datasets, but there is no evaluation on multiple

[1] https://github.com/cimejia/novel-FER-datasets.git

[2] https://github.com/muxspace/facial_expressions

[3] https://ispl.korea.ac.kr

datasets. In our case, we generate a merged, balanced and unbiased test dataset for the evaluation of distinct models trained on different datasets. Finally, we emphasize that the commonly employed technique for balancing datasets is data augmentation, unlike our strategy based on a state-of-the-art generative model.

B. ARTIFICIAL FACIAL DATASETS

Facial expression research has led to a wide range of applications based on deep learning algorithms, which require large volumes of data as a training resource. Collecting and labeling these datasets is costly and error-prone. Also, no control over features makes the distribution of classes unbalanced and biased, and facial information requires the consent of the individual due to privacy and misuse [34], [35]. These drawbacks reduce the quality of real datasets and affect the learning of models. Hence, artificial facial datasets emerge as a very promising alternative [36]. Instead of collecting, images are generated by computer, which reduces time and effort. This technique is scalable as needed, the labeling is automatic and more reliable, the user can control the process according to specified parameters, and privacy problems are minimized by not dealing with real people. A key utility for our work is to balance the rare cases as there are less frequent categories of gender, age or ethnicity. In summary, artificial facial datasets tend to be less noisy and more controlled than in-the-wild datasets, which is of potential benefit for improving emotion recognition. The current state-of-the-art in the generation of artificial facial images is advanced and in constant development. Several methods allow the generation of highly realistic facial images and control of features. They can be divided into two groups: geometric methods based on images and those using image synthesis techniques.

1) GEOMETRIC METHODS

The classic *data augmentation* applies geometric transformations such as rotation, horizontal and vertical flips, translation, reflection, scaling and random clipping. These effects can be achieved in a customized manner using conventional graphic libraries, however, the automatic tools provided by TensorFlow and Keras are preferred. Here, there is no need to train a deep neural network on a facial dataset, but the generated images are only slightly modified copies of the original images. There are also 2D and 3D modeling techniques that require reference images, scans and geometric parameters. These techniques allow precise control over the position, expression and lighting. For instance, Jin et al. [37] use FACSGen, a 3D face program, based on facial action coding system, to generate a synthetic dataset consisting of 1000 individuals, each one having 7 expressions. This dataset is used as unlabeled data and combined with the labeled data as input for a deep neural network. The training incorporates association learning, which rewards or penalizes an association based on the similarity between the labeled and unlabeled feature. Experimental results with RaFD and Oulu-CASIA datasets show that the discrimination capability of the deep network improves on the same dataset. Gao et al. [34] explore the use of synthetic datasets for face alignment. Different faces are created from various facial angles and attitudes, and adjusting gender, ethnicity and expression using a 3D face-generating middleware model (FaceGen[4]) Results suggest that adding synthetic datasets to the real ones to train the face alignment network can improve accuracy. Facial image generators can simulate variations in pose and facial expression, thus reducing the biases of real-world datasets. Vonikakis et al. [38] address the scarcity of dimensional emotion datasets. Morphing between facial images of categorical expressions is used to generate synthetic images that can be mapped to valence and arousal space with full control of the distribution and automatic dimensional labeling. The "MorphSet" dataset is presented with 167 individuals and approximately 342 expressions per individual, giving a total of 57K+ images. A ResNet-18 model is trained on MorphSet, AffectNet and Aff-Wild to predict valence and arousal, and evaluated on the AffectNet validation subset and a 20% of unseen Aff-Wild and MorphSet images. Performance metrics favor MorphSet, suggesting that it is suitable for supervised learning of dimensional emotions. Kollias et al. [23] use a facial image with a neutral expression to generate a new one, but with a different categorical or dimensional expression. The process performs face detection and landmark localization, fits a 3D morphable model, deforms the reconstructed face, adds the desired expression and blends the new face with the original image. The resulting facial images augment the data and train deep neural networks on several dimensional or categorical datasets, verifying better emotion recognition. Wood et al. [39] use a parametric 3D face model. The synthetic faces are rendered with an extensive library of high quality artistic resources. A generative model produces a 3D facial template to which the following elements are added, each one randomly selected: identity, expression, texture, hair, clothing and environment. Consistent and automatic labeling and full control of the variation and diversity of the dataset are highlighted. A dataset of 100K realistic and expressive synthetic human faces is provided[5] and evaluated on real datasets in two tasks: face parsing and landmark localization. The results demonstrate that models trained with synthetic data can generalize to real datasets, which could lead to other face-related tasks. Based on this work, Bae et al. [40] create DigiFace-1M,[6] a large-scale dataset of more than one million synthetic face images to avoid ethical issues, labeling noise and data bias. The computer graphics pipeline is based on head scans of a small number of people with consent. Correctness of labels is guaranteed and full control of data distribution ensures a fair dataset. Accuracy comparable to GAN models trained with millions of images of real faces is achieved.

[4] http://www.facegen.com
[5] https://github.com/microsoft/FaceSynthetics
[6] https://github.com/microsoft/DigiFace1M

2) IMAGE SYNTHESIS

This is a very effective method to artificially generate images containing a desired photorealistic facial expression. One of the most widely techniques is the use of generative adversarial networks (GANs), which can be trained on real datasets to generate images that form synthetic face datasets [41]. Since the original GAN could not generate facial images with a specific facial expression referring to a specific person, some methods conditioned on expression categories have been proposed. Colbois et al. [42] use StyleGAN2 and semantic editing of the latent space to generate facial images with controlled variations of pose, illumination and expression. The synthetic dataset is ''Syn-Multi-PIE'' and imitates the Multi-PIE[7] dataset. The experimentation indicates that the identities generated are new and meet privacy and accuracy requirements. The evaluation suggests that the synthetic dataset could replace the real one and obtain similar conclusions about face recognition performance. Boutros et al. [35] created ''SFace'', a privacy-friendly synthetic face dataset using StyleGAN2-ADA trained on the CASIA-WebFace dataset. The training is conditioned by assigning identity labels as class labels. Associating one real identity to another with the same class label of the synthetic dataset is hardly possible. The evaluation suggests that the use of SFace to train face recognition models can achieve high verification performance. Bozorgtabar et al. [43] synthesize photorealistic facial images conditioned by facial expression. An encoder-decoder uses the shared latent representation between image domains and a face landmark heatmap as a representation of facial expression. The Oulu-CASIA VIS dataset is extended with the synthetic images to train an expression classifier that achieves an average accuracy higher than the state-of-the-art. The synthetic images are of higher quality compared to IcGAN and CycleGAN, and convenient for data augmentation and FER performance improvement. Deng et al. [44] present ''DiscoFaceGAN'', a model for generating realistic facial images of virtual people with a precise control of target face properties. The method is based on an interpretable and highly disentangled latent representation for pose, expression and illumination. The training incorporates 3D priors and an imitative-contrastive learning framework. This allows Qiu et al. [45] to generate the Syn_10K_50 and Syn-LFW synthetic datasets used to train and test a face recognition model, respectively. For real face datasets, the authors use CASIA-WebFace[8] for training and LFW[9] for testing. By using identity and domain mixup, the cross-dataset evaluation shows the great potential of synthetic data for face recognition. Karras et al. [31] from NVIDIA propose StyleGAN2 to fix the defects and improve the quality obtained with StyleGAN [46]. Material related to both projects is available on GitHub.[10][11] A large set of generated images using FFHQ[12] (Flickr-Faces-HQ) at 1024×1024 as training dataset can be found on Google Drive.[13][14] A new StyleGAN3 version [47] has been developed for animation and video. Some example face images are available.[15] Websites such as this-person-does-not-exist.com and unrealperson.com implement GAN to download realistic faces of non-existent people via scripts. An example is the dataset presented in [48], which contains more than 5K images manually labeled into female and male classes, but the author warns of possible misclassification. Finally, Esser et al. [49] propose VQGAN, an architecture that combines the convolutional efficiency of the local interactions defined by the kernels with the global expressiveness of the long-range interactions of transformers. An autoencoder learns a compressed semantic information of the image (codebook) to compress and reconstruct an image with adversarial training using a discriminator. A transformer model learns to generate a sequence of this codebook just like an autoregressive model, but to perform high-resolution image synthesis.

Image synthesis using GAN and its variants have achieved state-of-the-art results [37], [43]. However, synthetic image generation inherits the problems of real datasets, so fully representing the complexity and diversity of facial expressions in the real world is a problem to be solved. We hypothesize that real datasets should first be perfected and then generate synthetic data to replace them completely. It is still necessary to train and evaluate the models with real images to ensure generalization in practical applications [37]. However, the complementarity of synthetic data to improve FER performance is supported by some studies reviewed in this section. The creation of synthetic facial images is useful mainly to balance our evaluator dataset. We present the use of a GAN, but the difficulties of training, the lack of control over the process and the large amount of time required are important drawbacks. In addition, the results are not entirely satisfactory in terms of the quality, expressiveness and realism. This is due mainly to the training datasets.

Recent advances in generative AI, mainly focused on lack of control, higher quality and ease of use [50], provide free and open source tools that can be downloaded and used directly without training. Stable Diffusion is a text-to-image model for generating high quality digital images from natural language descriptions. We know a pair of references where this model is used to generate face images. First, Beniaguev, D. [51] presents the SFHQ (synthetic faces high quality) dataset,[16] which contains 425K high quality 1024×1024 and

[7] https://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html
[8] https://www.kaggle.com/datasets/ntl0601/casia-webface
[9] http://vis-www.cs.umass.edu/lfw/
[10] https://github.com/NVlabs/stylegan
[11] https://github.com/NVlabs/stylegan2
[12] https://github.com/NVlabs/ffhq-dataset
[13] https://drive.google.com/drive/folders/1mTeo3J3Jo6aYImBshLM6XRl_Ua8fqgVW
[14] https://drive.google.com/drive/folders/14lm8VRN1pr4g_KVe6_LvyDX1PObst6d4
[15] https://github.com/NVlabs/stylegan3
[16] https://www.kaggle.com/datasets/selfishgene/synthetic-faces-high-quality-sfhq-part-4

curated face images. The original inspirational images come from datasets of paintings, drawings, 3D models, as well as images generated by Stable Diffusion (v1.4 and v2.1) using various prompts that cover a wide range of identity, ethnicity, age, pose, expression, lighting conditions, hairstyle, hair color and facial hair. The process that ''brings to life'' face-like images consists of each original image is represented by encoder4editing (e4e) in the latent space of StyleGAN2 and then verified to be photorealistic and of high quality. Facial landmarks and face parsing semantic segmentation maps are provided in the dataset. Second, a face dataset divided into training, validation and invalid folders is available in Kaggle [52]. There are about 2500 512 × 512 portraits of males and females generated from simple prompts using Stable Diffusion v1.4 in an Azure VM. Both works mention frequent errors in generation, so some method of correction is needed to create a useful dataset. We design structured and detailed prompts to match the images to the desired result, make an exhaustive selection of the correct images and automatically add the emotion category labels.

III. MATERIALS AND METHODS

This work is intended as a contribution to improve the recognition of emotions associated with facial expressions in real scenarios. Typically, research efforts are directed towards finding the ideal model, but first it is necessary to work on the quality of the data. The models are highly dependent on the training dataset, so they also learn of its shortcomings. We focus on the major problems of in-the-wild FER datasets in terms of quantity, diversity and generalization. Our proposal consists of the following aspects: (1) the creation of a large dataset by merging several known in-the-wild datasets to increase the variety of individuals and facial images; (2) this new dataset is used to train a more generic FER model, for which it is divided into training and test subsets. The latter is called the evaluator dataset, which is designed to be a benchmark of the generalization capacity of the models trained on different datasets. In addition to combined, the evaluator dataset is balanced, unbiased and well labeled. These conditions are met by the equal distribution of gender, age and ethnicity, the generation of synthetic image, and the exhaustive verification of facial expression images; (3) the creation of the first and largest artificial FER dataset using Stable Diffusion. This product may be useful for research and development in emotion recognition; (4) the selection of the most suitable convolutional networks to be trained and evaluated with the datasets considered in this work. These stages are described in this section, whereas the experimental part, which includes training (5) and evaluation using the single- and cross-dataset approaches (6), is explained in the next section.

A. CREATION OF THE MERGED DATASET

Because data collection is a costly task, we address the problem of lack of generality by combining existing datasets to increase the number of different individuals and cover a wider range of faces and expression variations. We consider three in-the-wild datasets, two of which are very popular and researched in the field of FER such as FER2013[17] and AffectNet,[18] whereas NHFI[19] (natural human face images) is more recent and less known, but is designed to provide images with better manual annotation. Factors such as availability, size, image format and emotion categories make these datasets convenient for our work. However, they also exhibit problems commonly mentioned in FER. As a consequence of the subjectivity of emotion perception, mislabeling or misclassification is one of the main causes affecting the performance of recognition models. In a previous work [26], we developed a method to reduce this problem by refining the dataset with several successive filterings. This process is applied to each of the datasets considered here, resulting in a new reclassified version. Table 1 presents the emotion categories and the number of facial images in total and per-category for each original dataset and its reclassified version.

TABLE 1. Distribution of emotion categories and number of facial images of the datasets considered.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
FER2013	4953	547	5121	8988	6198	6077	4002	35886
NHFI	890	439	570	1406	524	746	775	5350
AffectNet**	4300	4300	4300	4300	4300	4300	4300	30100
Total	10143	5286	9991	14694	11022	11123	9077	71336
FER2013*	4817	532	3842	9202	7074	6090	4329	35886
NHFI*	336	514	383	1585	1042	962	528	5350
AffectNet*	4394	3893	5004	4594	4224	4071	3920	30100
Merged	9547	4939	9229	15381	12340	11123	8777	71336

* reclassified
** balanced

TABLE 2. Properties of the FER datasets.

Dataset	Images	Resolution	Color mode	Format
FER2013	35886	48x48 px.	Grayscale	JPG
NHFI	5350	224x224 px.	Grayscale	PNG
AffectNet	30100	224x224 px.	RGB	PNG

The reclassified versions present higher reliability and lower intraclass variability, so the models trained on these reclassified versions achieve state-of-the-art recognition rates for FER2013, NHFI and AffectNet. These improved versions of the datasets are used to form the merged FER dataset. Although the original AffectNet dataset contains 287,401 images, we apply *downsampling* considering the category with the lowest number of images (disgust) to obtain a reduced, but balanced version of 4,300 randomly selected images per category. Thus we avoid the huge imbalance that characterizes this dataset, which could cause a bias in the training of a model.

It is important to emphasize that each dataset is distinguished by the number and properties of its facial images. Therefore, the combination of the datasets involves the mixture of different attributes, as shown in Table 2. This diversity complicates training and evaluation, but is essential to know the generalization capability of the models and datasets considered.

[17] www.kaggle.com/datasets/deadskull7/fer2013
[18] mohammadmahoor.com/affectnet-request-form/
[19] www.kaggle.com/datasets/sudarshanvaidya/random-images-for-face-emotion-recognition

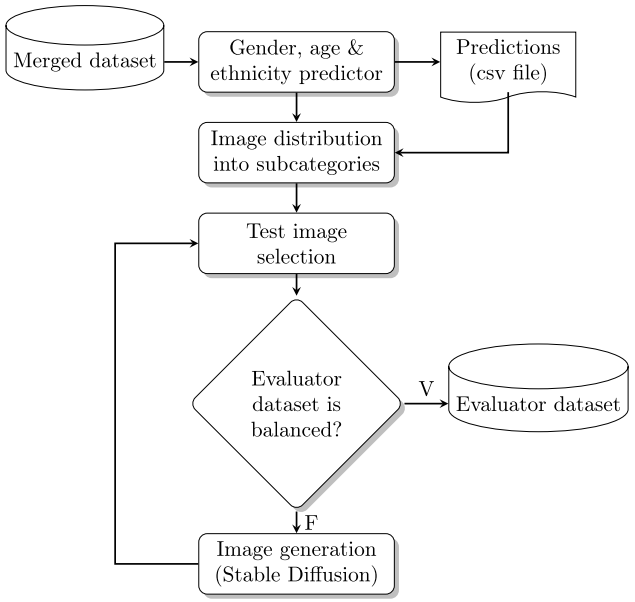


FIGURE 1. Methodology to create the combined, balanced and unbiased evaluator dataset.

A structure based on directory and subdirectories is created to organize the images of the reclassified versions of FER2013, NHFI and AffectNet in a single dataset. Each subdirectory corresponds to a category of emotion by means of a numerical label and in alphabetical order (0: angry, 1: disgust, 2: fear, 3: happy, 4: neutral, 5: sad and 6: surprise). This is very useful in the implementation phase to optimize the code and infer the list of classes automatically. As we work with the improved versions of the datasets, the image files of each dataset are moved directly into the respective categories. Previously, we verify the absence of duplicate file names to avoid overwriting and modification of the total number of files. As a result, a larger and more diverse dataset is available for the FER domain.

B. CREATION OF THE EVALUATOR DATASET

A model trained on the merged dataset reduces dependence on a specific dataset and extends its generalization to more datasets. This model can achieve better performance in emotion recognition in real-world environments due to the wider variety of data. For this purpose, the merged dataset is divided into training and test subsets. The latter is designed to be a evaluator dataset to measure and compare the performance and generalization of emotion recognition models. In addition of being mixed, a good benchmark must be balanced and unbiased, which are key properties for evaluating how good and generic an emotion recognition model is. The creation of the evaluator dataset with these requisites follows the working methodology shown in Figure 1.

1) GENDER, AGE AND ETHNICITY PREDICTION

More than large, the evaluator dataset should be representative of the general population, i.e., all groups of people have to be considered. Among the most relevant attributes for forming groups of people for social and commercial purposes are gender, age and ethnicity [53]. These attributes can be analyzed in a facial image and are often responsible for biases in FER datasets. Therefore, we propose a fair dataset containing the same number of facial images for all gender, age and ethnicity groups. Precisely, automated recognition of these features has become one of the most recognized fields of deep learning for its applications in social networks, video surveillance, biometric analysis and other uses [54].

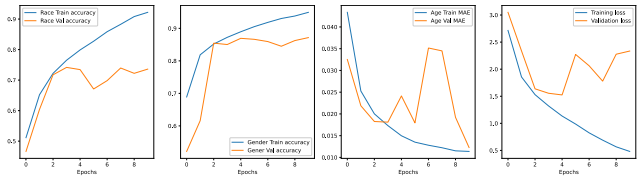


FIGURE 2. Learning curves of the age, gender and ethnicity prediction.

We leverage a gender, age and ethnicity predictor available at GitHub.[20] The author uses a convolutional network trained on the UTKFace[21] dataset, which has 20k+ in-the-wild facial images, containing a single face in each image, provides the aligned and cropped facial images, with their respective age, gender and ethnicity labels. Although the CNN implementation is available, the resulting model weights are not in a format suitable for reuse. We proceed to train the convolutional network by running the Python script[22] published in the same repository, which we modify slightly to save the model in h5 format,[23] obtaining the performance plots shown in Figure 2.

The learning curves behave acceptably. For the categorical variables of gender and ethnicity, training and validation accuracy curves are presented. Despite a separation, both reach a significant height, indicating good accuracy. For the quantitative variable of age, the mean absolute error (MAE) metric is plotted, whose curve is very close to zero in the last epoch, both for training and validation. We perform several predictions using face images from the UTKFace dataset, the results are presented in Figure 3. At the top of each image are the prediction values, whereas at the bottom are the truth labels. There is a good approximation in most cases. This model accuracy is sufficient to avoid us an exhaustive manual review of all images in the three datasets considered.

We use this model to predict gender, age and ethnicity for all the facial images. This facilitates the selection of the images that will form the evaluator dataset. Through a script,[24] the model in h5 format is loaded. All images are read to perform the feature prediction, obtaining the three values

[20] https://github.com/Sobika2531/Age-Gender-And-Race-Detection-Using-CNN
[21] https://susanqq.github.io/UTKFace/
[22] https://github.com/cimejia/novel-FER-datasets/blob/main/FER-evaluator-dataset/AGR-prediction/AGRprediction-V2.py
[23] https://github.com/cimejia/novel-FER-datasets/blob/main/FER-evaluator-dataset/AGR-prediction/best_model_agr.h5
[24] https://github.com/cimejia/novel-FER-datasets/blob/main/FER-evaluator-dataset/AGR-prediction/AGRprediction-test-V2.py

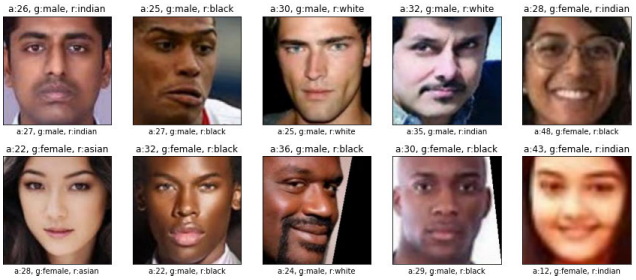


FIGURE 3. Results of the prediction of gender, age and ethnicity for some images from the UTKFace dataset.

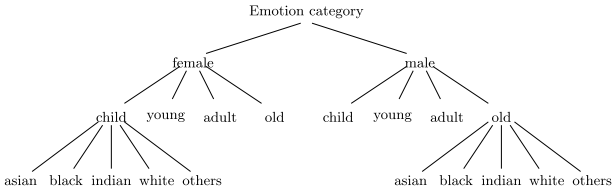


FIGURE 4. Generation of subcategories based on gender, age and ethnicity.

that are stored together with the image name in a csv file. This list is saved as a text file, so there are 7 files in total, one for each category. This procedure is applied for the reclassified version of FER2013, NHFI and AffectNet.

The prediction of age is a numerical value, unlike gender and ethnicity. To convert it into a qualitative value, we define the following rules: age less than 15, ''child'' class is assigned; age of 15 or more and less than 30, the class is ''young''; age of 30 or more and less than 65 is ''adult''; and age of 65 or more is the class ''old''.

In addition to simplifying the selection of facial images, the prediction of gender, age and ethnicity gives us an approximation of the imbalance of these variables in each dataset. This problem is analyzed and showed in Appendix B. Therefore, our evaluator dataset includes all defined gender, age and ethnicity groups equally, i.e., the same number of facial images within each class, which is explained as follows.

2) FACIAL IMAGE DISTRIBUTION

We organize the structure of folders and subfolders according to the hierarchy shown in Figure 4. One folder is dedicated to each of the seven categories of emotion, whereas the subfolders correspond to all possible combinations of gender, age and ethnicity. As a result, each type of emotion is composed of 40 subcategories ($2\times4\times5$), which ensures a balanced and unbiased distribution of images.

The nomenclature used for the subcategories is key to the distribution of the images. The name is formed by combining gender, age and ethnicity, e.g., the ''angry'' category has a folder named ''*female-adult-asian*'', which stores the facial images detected as adult women of Asian origin with an angry expression. All images of each dataset are automatically distributed within the 40 folders using a Python script[25] supported by the *csv* and *shutil* libraries. The prediction file is read as text format, whose lines contain the values of file name, gender, age and ethnicity delimited by a comma. Iteratively, each folder name is formed with these values. Finally, we move each image file to its respective folder. This process is performed for each emotion category, obtaining a new version of the dataset structured under 7 categories and 40 subcategories, one for each combination of gender, age and ethnicity. This automated distribution considerably accelerates the equal and unbiased selection of images, in contrast to a manual and exhaustive selection.

TABLE 3. Distribution of training, validation and test subsets for each dataset.

Dataset	Train subset (80%)	Validation subset (10%)	Test subset (10%)
FER2013 (reclassified)	28708	3589	3589
NHFI (reclassified)	4280	535	535
AffectNet (reclassified)	24080	3010	3010

TABLE 4. Number of images selected per dataset and emotion category for the test subset.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
FER2013	67	50	70	80	80	77	64	488
NHFI	62	62	59	79	75	69	55	551
AffectNet	77	70	70	80	75	72	75	519
Total	206	182	199	239	230	218	194	1468

3) TEST FACIAL IMAGE SELECTION

First, it is necessary to determine the number of images to be selected from each subcategory of gender, age and ethnicity and for each category of emotion. This number results by dividing the size of the evaluator dataset by the 40 subcategories. The total of images is conditioned by the smallest test subset of the three datasets considered for combination. Table 3 shows the division of the datasets into three subsets: *train*, *validation* and *test*, applying 80%, 10% and 10%, respectively. This proportion is one of the most recommended for machine learning. Consequently, the size of the NHFI test subset (535) determines the number of images that should be selected from each dataset.

This quantity is not perfectly divisible by the total number of subcategories. Selecting one image for each of the 40 subcategories and for each of the 7 categories of emotion, we obtain 280 (40×7) images in total, which is equivalent to 5.23% of the dataset, i.e., lower than the estimate. Selecting 3, the result is 840 (3×40x7), which is 15.7%, i.e., higher than the estimate. Therefore, we select by visual inspection 2 facial images for each of the 40 subcategories, a total of 560 facial images corresponding to 10.47%, similar to the recommended percentage. Table 4 presents the number of facial images selected in total, per dataset and category within each dataset, which should be 1680, 560 and 80, respectively.

During the selection task, some subcategories had only one or no facial image. This occurs mostly in classes associated

[25]https://github.com/cimejia/novel-FER-datasets/blob/main/FER-evaluator-dataset/AGR-prediction/folders-distribution.py

TABLE 5. Number of remaining facial images to balance the evaluator dataset.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
FER2013	13	30	10	0	0	3	16	72
NHFI	18	18	21	1	5	11	25	99
AffectNet	3	10	10	0	5	8	5	41
Total	34	58	41	1	10	22	46	212

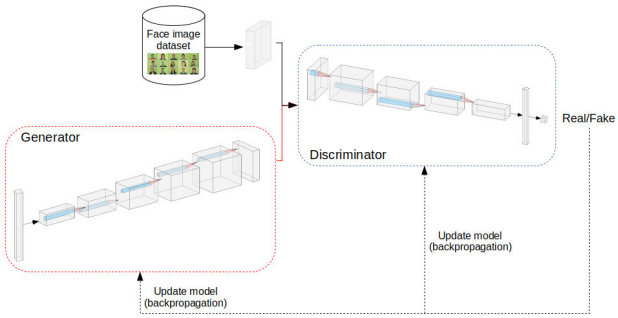


FIGURE 5. GAN architecture.

with children and old people, as well as Asians, Indians and other non-white and non-black people. Therefore, the FER datasets have an insufficient number of images to satisfy the gender, age and ethnicity criteria. This problematic avoids to complete the evaluator dataset, so we must generate the needed images to achieve the balance. Table 5 shows a detail of the number of facial images required for balancing, both in total and per dataset and emotion category.

4) ARTIFICIAL FACIAL IMAGE GENERATION

Synthetic image generation is our strategy to obtain the remaining 212 facial expression images. In this section, we present the use of GAN and Stable Diffusion, and compare the results to select the most appropriate technique.

a: GAN [55]

As reviewed in section II, synthetic images to augment facial datasets are generated especially with GANs. Usually this occurs for the training stage. In our case, it is required for the evaluation stage to achieve a perfectly balanced and unbiased test dataset. We reuse a designed GAN [56], but adapted it to our problem of interest.[26] This architecture combines a generator and a discriminator model into a single larger one (Figure 5).

The task of the generator is to produce synthetic (fake) images from random latent vectors, whereas the discriminator is in charge of deciding whether these images pass as real. The name of the entire network is due to the fact that both models work adversarially. Training the discriminator improves the differentiation of the images, the result is used to train the generator in order to obtain images progressively more similar to the original dataset. This process is performed until the discriminator assigns a synthetic image as real.

[26]https://github.com/cimejia/novel-FER-datasets/tree/main/FER-artificial-dataset/GAN

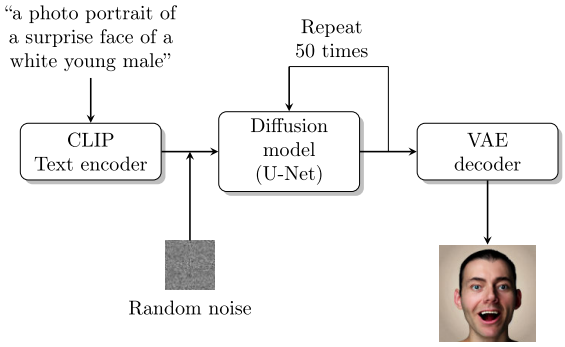


FIGURE 6. Stable Diffusion architecture.

After training the GAN using the facial images of the disgust category, a sample of the results obtained is shown in Figure 18 (Appendix C). Not all images are useful. Although the appearance is of a face, certain features have not been fully formed, so we should perform a selection process of the best samples. We emphasize that the only mechanism to control the expression of people in synthetic images is the content of the input dataset. Therefore, training and image selection tasks must be performed for each emotion category. This is a difficult and time-consuming process (hours or days depending on hardware capacity). In addition, the quality of the resulting synthetic images is conditioned by the input images, which are of low resolution.

b: STABLE DIFFUSION

The drawbacks detected in GAN motivate the search for an alternative to create artificial images. Recently, *generative IA* has developed rapidly and has achieved great relevance with impressive results. In particular, diffusion models have become the state-of-the-art in image synthesis and superresolution [29]. Among the most prominent generative tools, *Stable Diffusion* allows the generation of high-quality digital images from natural language descriptions. The images generated are similar to those used to train the model.

The architecture[27] consists of a VAE (variational autoencoder), a diffusion model based on a U-Net network and a CLIP (contrastive language image pretraining) text encoder. Each component with its own neural network. The operation can be explained in two stages: training and inference. Both stages use latent representations learned by the VAE to encode and decode the images. This representation is a compressed and probabilistic version that allows optimizing the process and generating variations of the original image. During the training, diffusion progressively adds noise to the input images to generate noisy versions. The original images and their noisy pairs are the dataset to train a U-Net model to map noisy to high-quality images. In inference (Figure 6), a random Gaussian noise and a textual description (prompt) are the inputs to the model. Before, the prompt is passed

[27]https://keras.io/examples/generative/random_walks_with_stable_diffusion/

through the CLIP encoder to generate the text embedding that conditions the visual content. The U-Net model is responsible for predicting and eliminating noise (denoising), generating the output based on the prompt using cross-attention layers. This process is repeated a specified number of times (50 by default), where the added noise is gradually reduced. Finally, the latent representation passes through the VAE decoder to obtain the final image.

Stable Diffusion is developed by *Stability AI* as open source and free. Through the company *Hugging Face*[28] are available the code and weights,[29] a demo Web version,[30] a programming notebook[31] and the *diffusers* library for download and installation. We use diffusers supported by the *Torch* library and import *StableDiffusionPipeline* to instantiate version 1.4 of the model. It is a type of diffusion model pretrained for vision and used as an inference tool. We do not retrain the model on the FER datasets, as it was originally trained with 512 × 512 images from a subset of LAION-5B,[32] a dataset of 5.85 billion image-text pairs [57]. The model runs in inference mode with a few lines of code[33] and a text sentence (*prompt*) as argument.

The generation of proper synthetic images depends on the quality of the prompt. It is a textual indication formed by several tokens interpreted by the AI to convert it into an image according to our needs. As this is not a trivial task, support tools are available, e.g., *Lexica*[34] is a search engine for images generated by Stable Diffusion that allows visualizing the prompt used for such images. After many experiments with this tool, we suggest the following structure to achieve good prompts.

$$\textit{Style} + \textit{main topic} + \textit{adjectives} + \textit{reinforcement} + \textit{technical data}$$

1) The style is the type of image desired, e.g., illustration, digital art, photography, canvas, cartoon, drawing, painting or portrait.
2) The main topic should be described specifically with the objects to be observed in the image. This refers to a landscape, close-up or general view and what should go in the center such as a person, animal, plant or anything.
3) One or more adjectives that precisely define the properties or attributes of objects.
4) Include words to express an action, complement or reinforce elements with synonymous or alternative terms.
5) Technical data about the image and the degree of detail to achieve more realistic images, e.g., high quality, 4K or hd.

The parts indicated must be well linked, converge in the same concept, without inconsistencies or contradictions and avoiding divergences. The order influences the importance that the model gives to these parts, separating them with a comma rather than with a long sentence that groups them all.

By using the suggested guidelines, we create facial images with a specific emotion that are closer to the desired result. Table 11 (Appendix D) shows the prompts used to produce the remaining images to balance our evaluator dataset. These images are obtained by replacing the word ''person'' with the phrase specifying gender, age and ethnicity. For instance, the *''female-old-asian''* subcategory of the angry category in the NHFI dataset had no associated facial images, so we generate them by specifying the phrase ''A detailed photographic portrait of a perfect face of an asian old female, feeling an extreme rage, expressing a very angry face, features well-defined, facing the camera, realistic, 4K, hd''.

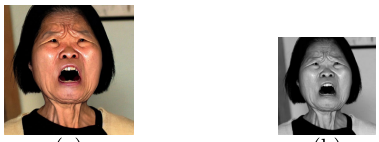

(a) (b)

FIGURE 7. Synthetic facial image for the angry category and the female-old-asian subcategory. (a) Generated image and (b) Image converted to the NHFI format.

Figure 7 shows the resulting image which has the default size of 512 × 512 pixels and RGB color. We convert to 224 × 224 grayscale (Figure 7b) according to NHFI. Thus 212 facial images are obtained to balance the evaluator dataset. However, it is necessary to generate a larger number of images because not all of them clearly show the type of emotion requested. For this reason, Table 11 includes the prompt effectiveness, some emotions being more complex than others. The happy and sad categories are the easiest to produce, 9 of 10 images are well generated. Next comes the neutral category, with 8 of 10, whereas the disgust and angry categories present the highest complexity and time as only 3 and 4 images of 10 are appropriate, respectively.

The images artificially generated to balance the test subset for the angry category from NHFI are shown in Figure 19 (Appendix C). Faces of female, Indian-origin, old age and children are particularly needed. Similarly, we generate the images according to gender, age and ethnicity for the test subsets of FER2013 and AffectNet. The three test subsets are merged to obtain a single balanced and unbiased dataset (Table 6), which is useful as a benchmark for evaluating the generalization capability of a recognition model in real-world applications.

C. CREATION OF THE ARTIFICIAL FER DATASET

The good performance of Stable Diffusion in terms of quality of synthetic images and generation time motivates us to

[28] https://huggingface.co/
[29] https://huggingface.co/CompVis/stable-diffusion
[30] https://huggingface.co/spaces/stabilityai/stable-diffusion
[31] https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/stable_diffusion.ipynb
[32] https://laion.ai/blog/laion-5b
[33] https://github.com/cimejia/novel-FER-datasets/tree/main/FER-artificial-dataset/Stable-Diffusion
[34] https://lexica.art/

TABLE 6. Distribution of the combined, balanced and unbiased evaluator dataset.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
FER2013	80	80	80	80	80	80	80	560
NHFI	80	80	80	80	80	80	80	560
AffectNet	80	80	80	80	80	80	80	560
Evaluator	240	240	240	240	240	240	240	1680

TABLE 7. Distribution of the artificial FER dataset generated with Stable Diffusion.

Subset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
Train artificial	550	707	968	1050	1239	1492	599	6605
Test artificial	240	240	240	240	240	240	240	1680
Total	790	947	1208	1290	1479	1732	839	8285

contribute with a fully artificial FER dataset. Labeling is automatic and controlled by the prompt, which determines the category. However, the balance and bias become difficult to control due to the variable effectiveness of the prompts. For the categories of disgust and surprise, it was necessary to include the corresponding action units in their prompts to improve the accuracy of the synthetic images. We use the diffusion model to artificially generate thousands of facial images considering all categories of emotion and the criteria of gender, age and ethnicity. Table 7 presents the distribution of the artificial FER dataset after the selection process of the best facial images.

The generation of each image involves the 50-step denoising process, which is relatively slow and memory consuming, taking about 11 seconds with the available hardware (Appendix E), i.e., 5 per minute and 300 per hour. Although the images are visually inspected to select the most suitable ones and avoid those that are cartoonish, too distorted, totally dark or with more than one person, the use of Stable Diffusion is much more efficient, simple to use and the images are of high resolution compared to GAN.

D. SELECTION OF THE MODEL

In previous sections, we have prepared single and merged FER datasets to train and evaluate emotion recognition models. This also includes the artificial dataset. The goal is to achieve an emotion recognition model with better generalization to real-world applications. To know whether the generalization improves, we perform training and evaluation of the deep learning models on each of the datasets analyzed here. This section describes the architectures of these models. A same architecture can perform different in function of the training dataset. This dependency means that the performance of a model is not necessarily the same for different datasets. Therefore, we identify the best model for each dataset.

The state-of-the-art tool for the problem of classifying facial images into emotion categories is the convolutional neural network (CNN). There are two main approaches: (1) a customized architecture, in which we create the layer structure and network hyperparameters from scratch, and (2) a transfer learning model, reusing an already created and pretrained public architecture, and adapting it to the problem

TABLE 8. Division of the datasets for the training process.

Dataset	TRAIN		TEST	Total
	Training (80%)	Validation (20%)		
FER2013	28321	7077	560	35958
NHFI	3915	974	560	5449
AffectNet	23669	5912	560	30141
Merged	55896	13972	1680	71548
Artificial	5284	1321	1680	8285

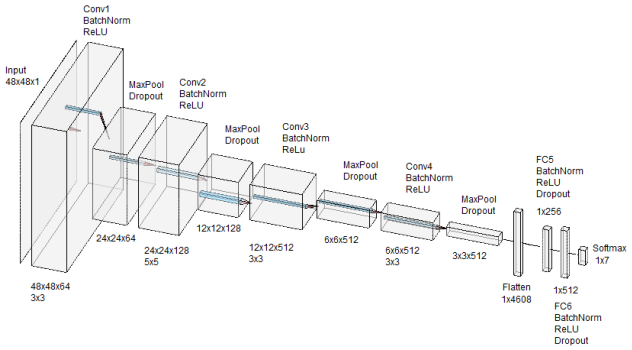


FIGURE 8. CNN architecture for the FER2013, AffectNet and merged datasets. We reuse the network designed by Akshit Bhalla [58].

of interest. Selecting the most suitable option is key to achieve the best results.

In previous work [26], several architectures with different hyperparameters are tested. A customized CNN presented the best performance for FER2013 and AffectNet, whereas the transfer learning technique, using CNNs based on EfficientNet and MobileNet, are the best for the NHFI and artificial datasets, respectively.

Figure 8 shows the architecture of the customized CNN, which combines a convolutional part and a classifier. The convolutional part receives the input image, which is processed by 4 convolutional layers that apply filters to extract features hierarchically, from the simplest to the most complex ones. Each convolution layer includes a relu activation function and maxpooling operation to reduce the dimensions of the feature maps. Regularization techniques such as normalization and dropout are used to optimize the network performance and reduce possible overfitting. The extracted features are passed in vector form to the classifier, which is an artificial neural network with two hidden layers and a softmax output layer that produces a probability value for each of the seven emotion categories.

Figure 9 presents the architecture of the network based on the transfer learning technique. The advantage is that the original image resolution of 224 $\times$ 224 is accepted as input, which is processed by the convolutional part of a pretrained model for feature extraction. We use the EfficientNet version B0 [59] and MobileNetV2 [60] for NHFI and the artificial dataset, respectively. These are state-of-the-art networks recognized at the time for their level of speed and optimization. The extracted features are received by the classifier in the form of a flattened vector. This is the input for a fully connected neural network with two dense layers of

TABLE 9. Training hyperparameters set for our experiments. These values are not determined by fixed rules, but are the result of several tests to find the most convenient ones.

Hyperparameter	FER2013	NHFI	AffectNet	Merged	Artificial
Input shape	48,48,1	224,224,3	48,48,3	48,48,3	224,224,3
Batch size	64	64	64	64	64
Pixel norm	1/255	No	1/255	1/255	1/255
Optimizer	Adam	Adam	Adam	Adam	Adam
Learning rate	0.01 to 0.00001	0.01 to 0.00001	0.0003 to 0.00001	0.0003 to 0.00001	0.01 to 0.00001
Loss function	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy
Metrics	loss & accuracy	loss & accuracy	loss & accuracy	loss & accuracy	loss & accuracy
Classes	7	7	7	7	7
Epochs	50	50	50	50	50
Augmentation	Yes	No	Yes	Yes	No

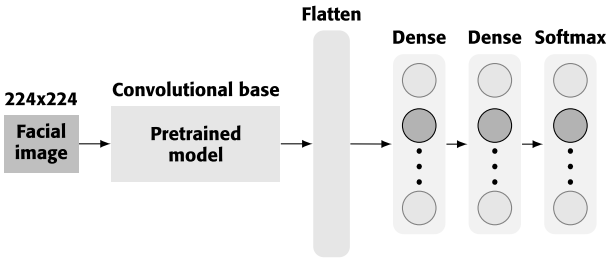


FIGURE 9. Pretrained CNN architecture for the NHFI and artificial datasets.

256 and 512 neurons, to which the relu activation function is applied, in addition to batch normalization and dropout regularization techniques to reduce potential overfitting. The softmax function of the last dense layer outputs a probability distribution corresponding to each of the seven categories of emotion to determine the class to which the input image belongs.

IV. EXPERIMENTS AND RESULTS

In this section, the experimental part and the results obtained are detailed. We explain how the datasets are organized for training and evaluation of deep learning models. To know the behavior of the model during training, we present and analyze the respective learning curves. Finally, we evaluate the performance using confusion matrices and an accuracy metric. The hardware and software platform is specified in Appendix E.

A. DATASETS

The fundamental resource for training is the dataset. Along the previous sections, we have prepared in total 5 datasets (Table 8), which are used for our experimentation. Each dataset is organized in two folders: ''TRAIN'' containing the facial images to fit the model, and ''TEST'' for the images of the balanced and unbiased test subset (only the artificial one is not unbiased), which allows to evaluate the model. At training time, the train set is automatically subdivided in a ratio of 80:20 into the training and validation subsets, respectively.

B. TRAINING

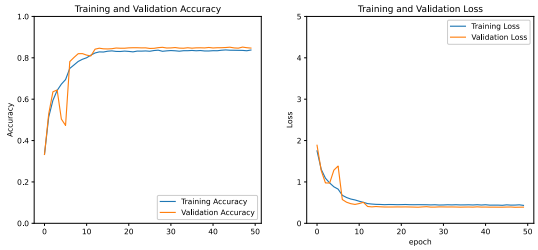
We perform one training on each dataset using the hyperparameters listed in Table 9 and the most proper network architecture, i.e., the customized CNN for the FER2013, AffectNet and merged datasets, whereas EfficienteNetB0 and MobileNetV2 for the NHFI and artificial datasets, respectively.

During training, the model learns to associate facial images with emotion labels. Using *ImageDataGenerator* from Keras, the facial images are passed to the model in batches of 64 images and automatically labeled with the respective category. This utility also handles data augmentation, pixel normalization and unification of the different resolutions and colors of the merged dataset. For each batch, the predicted and actual labels are compared using the *categorical_crossentropy* function, obtaining loss and accuracy for the training and validation subsets. The backpropagation and *Adam* algorithms (based on gradient descent) reduce the error by updating the weights as a function of the *learning rate*. This value decreases from an initial value to a minimum whether the loss does not improve after a few epochs. The loss and accuracy metrics for the 50 epochs are represented as learning curves with Matplotlib.

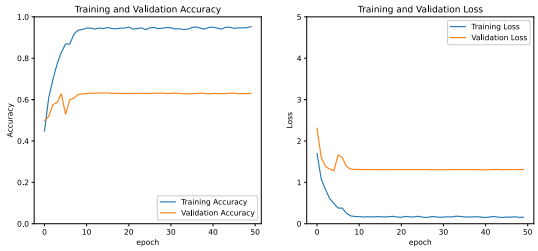
Figure 10 shows the learning curves of the convolutional network on each dataset. The accuracy (left) and loss (right) curves of the training and validation subsets are presented. Ideally, the accuracy curves should increase in height as the epochs advance (horizontal axis), whereas the error curves should approach zero. In addition, the training and validation curves should be very close to each other to avoid overfitting or underfitting. The best model performance is for the AffectNet, FER2013 and merged datasets in that order, whereas for the NHFI and artificial datasets overfitting is marked with a large separation of training and validation curves for both accuracy and loss. This indicates that the model fits the training data fairly well, but shows a regular performance on the validation images. We are particularly interested in the model trained on the merged dataset. The levels achieved by the accuracy curves are above 90% for training and 75% for validation, whereas the loss curve is very close to zero for training and below 1 for validation. This is an acceptable performance considering the size and variability of the dataset. Although there is a separation between the training and validation curves, the distance is not large. This suggests that the model may have good performance on the evaluator dataset.

C. EVALUATION

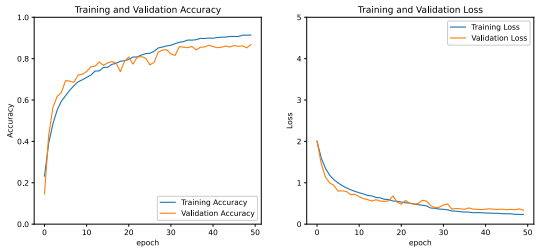
Practical use of a deep learning model requires not only knowing how good it is at training, but also how it performs



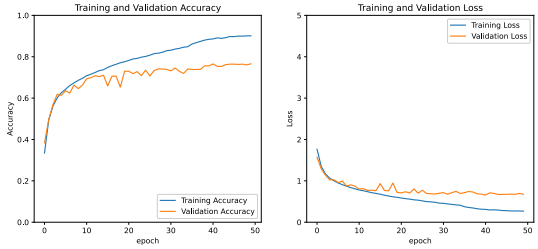
(a) Customized CNN trained on the FER2013 dataset.



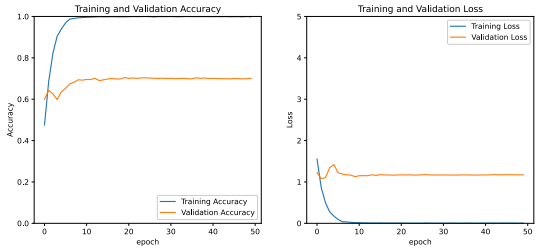
(b) EfficientNetB0-based CNN trained on the NHFI dataset.



(c) Customized CNN trained on the AffectNet dataset.

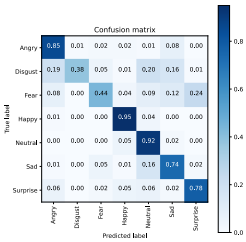


(d) Customized CNN trained on the merged dataset.

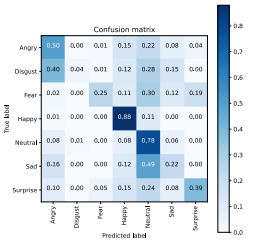


(e) MobileNetV2-based CNN trained on the artificial dataset.

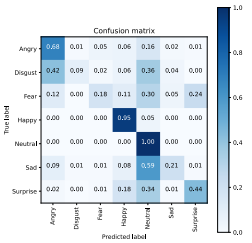
FIGURE 10. Learning curves of the training and validation stages of the convolutional networks on the datasets considered.



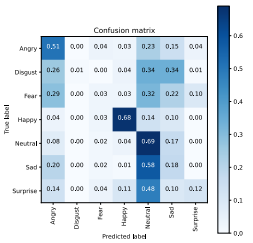
(a) Single-dataset: model trained and evaluated on FER2013.



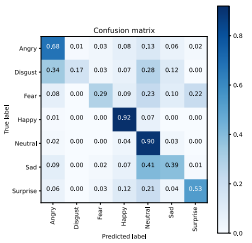
(b) Cross-dataset: model trained on FER2013 and evaluated on NHFI.



(c) Cross-dataset: model trained on FER2013 and evaluated on AffectNet.



(d) Cross-dataset: model trained on FER2013 and evaluated on artificial dataset.

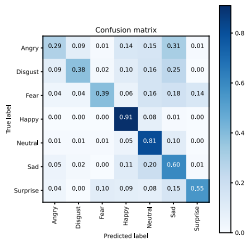


(e) Cross-dataset: model trained on FER2013 and evaluated on merged dataset.

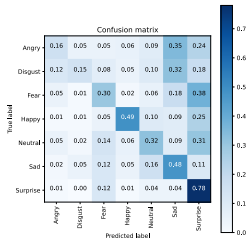
FIGURE 11. Single- and cross-dataset confusion matrices for the customized CNN model trained on FER2013 and evaluated on each test subset.

with new images. We conduct single- and cross-dataset evaluations to measure performance on the same dataset and also on the rest to determine the generalization capability. To this end, we use the models obtained in the training stage and the FER2103, NHFI, AffectNet, merged and artificial test subsets. It is essential that the facial images of the test subset be hidden from the model until evaluation. In single-dataset approach, each model is evaluated with the test subset of the same training dataset. Performance is specific and limited to a particular dataset, which is not very useful for diverse contexts as the real world. In cross-dataset approach, the test dataset is different from the training dataset. Typically, a single test dataset is used, but we extend the evaluation to more datasets that work as one. Our focus is on performance using the evaluator dataset, which provides a more general measure for practical applications. The results are presented visually and quantitatively through confusion matrices and the accuracy metric summarized in a comparative table.

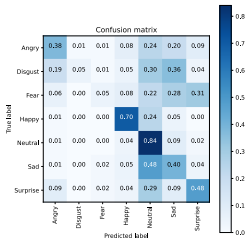
The *confusion matrix* shows the number of true positive, true negative, false positive and false negative predictions made by a classification model. In this case, we use the normalized values between 0 and 1. Figures 11, 12, 13, 14 and 15 are the confusion matrices for the FER2013,



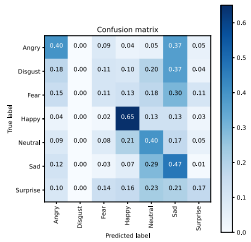
(a) Single-dataset: model trained and evaluated on NHFI.



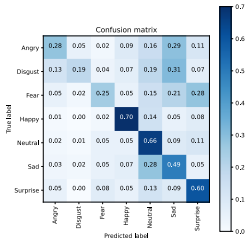
(b) Cross-dataset: model trained on NHFI and evaluated on FER2013.



(c) Cross-dataset: model trained on NHFI and evaluated on AffectNet.

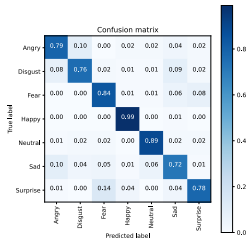


(d) Cross-dataset: model trained on NHFI and evaluated on artificial dataset.

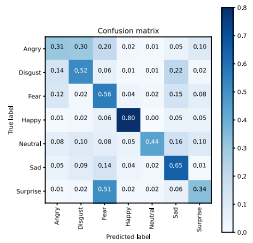


(e) Cross-dataset: model trained on NHFI and evaluated on merged dataset.

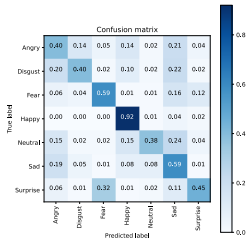
FIGURE 12. Single- and cross-dataset confusion matrices for the EfficienteNetB0-based model trained on NHFI and evaluated on each test subset.



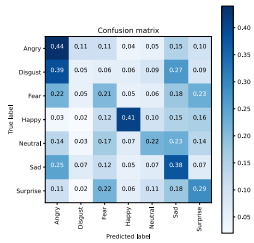
(a) Single-dataset: model trained and evaluated on AffectNet.



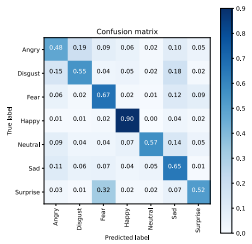
(b) Cross-dataset: model trained on AffectNet and evaluated on FER2013.



(c) Cross-dataset: model trained on AffectNet and evaluated on NHFI.



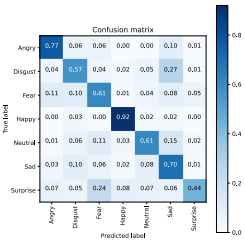
(d) Cross-dataset: model trained on AffectNet and evaluated on artificial dataset.



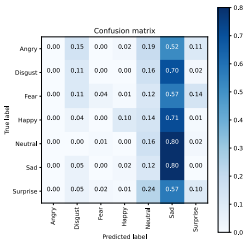
(e) Cross-dataset: model trained on AffectNet and evaluated on merged dataset.

FIGURE 13. Single- and cross-dataset confusion matrices for the customized CNN model trained on AffectNet and evaluated on each test subset.

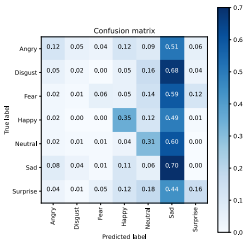
NHFI, AffectNet, merged and artificial test subsets, respectively. Each figure contains one single-dataset matrix and four cross-dataset matrices as there are five test subsets in total. All matrices use heatmap mode, where each cell is colored as a function of its value, i.e., the higher the cell value or the number of predictions, the greater the color intensity. The matrix is accompanied by a scale showing the range of colors and values. An ideal model performance would highlight the main diagonal of the matrix with the highest color intensity, whereas the remain cells should show the lowest color intensity. For all datasets, the confusion matrix that most closely approximates this behavior is the single-dataset approach (Figures 11a, 12a, 13a, 14a and 15a), so the recognition performance is expected to be acceptable for images from the same dataset. A similar situation occurs with the model trained on the merged dataset and tested with FER2013 and AffectNet (Figures 15b and 15d), which are not strictly cross-dataset cases, as the test images come from the largest datasets used for the merging. In contrast, the worst performance, reflected by a chaotic distribution of colors, is for the evaluation on the artificial test subset (Figure 11d, 12d, 13d and 15e). This suggests a significant dissimilarity between the real datasets and the synthetic dataset. The same happens in the opposite direction, i.e., when the model trained with the artificial dataset is evaluated on test subsets of real datasets (Figures 14b, 14c and 14d). The single- and cross-dataset evaluations of the model trained on the merged dataset show the best overall performance (Figure 15). Except for the confusion matrix of the artificial subset (Figure 15e), the others are very similar, displaying the main diagonal with the darker colors and the rest of the cells with the lighter colors. The merged dataset is the only case



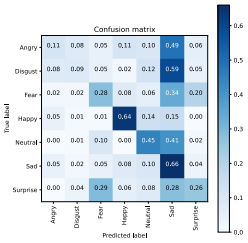
(a) Single-dataset: model trained and evaluated on artificial dataset.



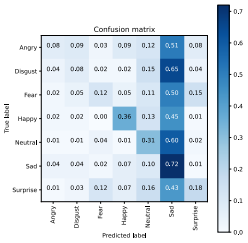
(b) Cross-dataset: model trained on artificial dataset and evaluated on FER2013.



(c) Cross-dataset: model trained on artificial dataset and evaluated on NHFI.

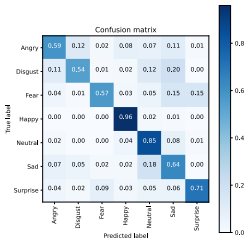


(d) Cross-dataset: model trained on artificial dataset and evaluated on AffectNet.

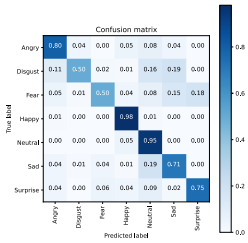


(e) Cross-dataset: model trained on artificial dataset and evaluated on merged dataset.

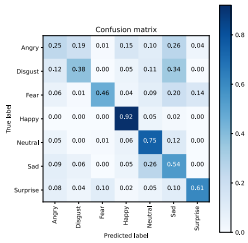
FIGURE 14. Single- and cross-dataset confusion matrices for the MobileNetV2-based model trained on the artificial dataset and evaluated on each test subset.



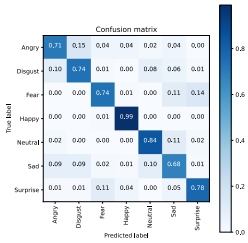
(a) Single-dataset: model trained and evaluated on merged dataset.



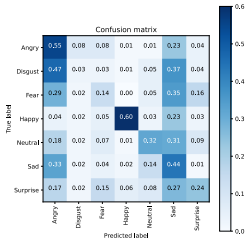
(b) Cross-dataset: model trained on merged dataset and evaluated on FER2013.



(c) Cross-dataset: model trained on merged dataset and evaluated on NHFI.



(d) Cross-dataset: model trained on merged dataset and evaluated on AffectNet.



(e) Cross-dataset: model trained on merged dataset and evaluated on artificial dataset.

FIGURE 15. Single- and cross-dataset confusion matrices for the customized CNN model trained on the merged dataset and evaluated on each test subset.

able to obtain homogeneous performance with itself and with the others.

The color gradient is a useful visual tool to quickly have an idea of the overall model performance and to identify patterns or trends in the data. It also informs us about the accuracy in each class individually. The best recognition in the real datasets corresponds mainly to the happy and neutral categories, with rare exceptions, whereas the disgust and fear categories are the worst recognized. Although the artificial dataset has acceptable single-dataset performance, cross-dataset evaluations indicate a remarkable bias for the sad category, which is explained by training the model with a larger number of images from this category.

For a quantitative evaluation, there are many values in the confusion matrix from the 7 emotion categories, which can be confusing. Therefore, an indicator that measures the overall quality of each model and dataset is needed to facilitate comparison. We use *accuracy* as a performance metric, which is calculated by summing the number of correct instances in all classes and dividing it by the total number of instances in the dataset. Table 10 summarizes the accuracy results obtained by each model on each test subset. This table is useful to compare the model performance with the dataset itself and with the rest. For instance, 0.7214 is the single-dataset accuracy achieved by the customized CNN model for FER2013, whereas 0.5179 is the cross-dataset accuracy achieved by the customized CNN model for FER2013, but trained on AffectNet. Therefore, different accuracy is obtained for the same dataset depending on the training dataset, even though the model architecture is the same.

Several results emerge from this comparative table.

- The main diagonal (orange) contains the accuracy values achieved on the same dataset. This single-dataset accuracy outperforms the cross-dataset accuracy for each

TABLE 10. Summary of the accuracies obtained for the single-dataset and cross-dataset experiments.

		Test subset				
Best model	Train subset	FER2013	NHFI	AffectNet	Artificial	Merged
CNN customized	FER2013	0.7214	0.4357	0.5054	0.3185	0.5542
EfficientNetB0	NHFI	0.3821	0.5607	0.4125	0.3143	0.4518
CNN customized	AffectNet	0.5179	0.5321	0.8125	0.2845	0.6190
MobileNetV2	Artificial	0.1875	0.2482	0.3554	0.6613	0.2637
CNN customized	Merged	0.7411	0.5589	0.7804	0.3310	0.6935

model and training dataset. This corroborates the strong dependence of the model on the dataset used for training.

- With the exception of NHFI, the rest of the datasets show good to very good single-dataset performance, especially AffectNet and FER2013. However, this can give a false sense of recognition peformance. It is not an indicator of how the model will perform with images that do not come from the training dataset. This accuracy may be acceptable for test images of the dataset itself, but when the model is tested with different datasets, the results are very low, indicating lack of generalization.
- The last column (green) shows the accuracy values provided by the evaluator dataset, which covers test images and features from different datasets. This is a more robust and generic metric for comparing models and datasets, as it indicates how good the model is for generalization to more datasets. This validates the hypothesis RQ2, so we propose this benchmark for a more general state-of-the-art in the emotion recognition task.
- Based on the proposed metric, we can expect better recognition results in real-world situations using the personalized CNN model trained on the merged dataset, which achieves the highest accuracy (red) and outperforms FER2013, NHFI and AffectNet in generalization capability by 13.93%, 24.17% and 7.45%, respectively. This validates the hypothesis RQ1, which states that combining several in-the-wild datasets to train a convolutional network allows obtaining a model with improved generalization, less dependent on a specific dataset and useful for practical applications.
- Cross-dataset accuracy considering only one test dataset (cells without color) is not a good measure of generalization, as it extends to a single dataset different from the training dataset. As expected, these accuracies are very low, suggesting the significant disparity of images and properties between the FER datasets. For instance, the customized CNN model is trained on FER2013 for JPG images of dimensions (48,48,1), but when evaluated on NHFI, the images are PNG of dimensions (224,224,3). Although the ImageDataGenerator utility performs the format conversion, the underlying dissimilarity degrades performance.
- Analyzing the cells without color by rows, the values can reflect the similarity between datasets. For instance, FER2013 is more similar to AffectNet, while AffectNet is more similar to NHFI. In contrast, all real datasets are less similar to the artificial dataset. This may be due to the low intraclass variation in synthetic face images and the domain gap between synthetic and real face datasets. Therefore, real facial image datasets cannot be completely replaced by synthetic ones, but it is possible to complement them for face-related tasks.

V. CONCLUSION

A proposal to improve emotion recognition is presented in this paper. The central idea is that training and evaluating models with merged datasets more closely approximates real-world scenarios. Our method addresses three fundamental aspects: (1) generate a large and more diverse categorical FER dataset created by combining in-the-wild datasets, rather than manual collection and labeling from scratch, which is time-consuming, labor-intensive and error-prone. We contribute a large dataset of 71,548 facial images, mixed in resolution, color, background, illumination and file format. To our knowledge, this is the first and largest merged dataset operating as one in FER, divided into training and testing subsets. It is useful for training an emotion recognition model with better generalization in real environments. (2) The test subset is designed to be unbiased and balanced in response to the need for a benchmark for a more realistic evaluation of the generalization of FER models and datasets. To this end, the organization of the evaluator dataset is based on representative characteristics of the population and the equal selection of facial images according to the different categories of gender, age and ethnicity of individuals. The lack of facial images for certain categories is addressed by Stable Diffusion, an effective tool generating the synthetic images to balance the dataset. As a result, this evaluator dataset has 1,680 images of facial expressions intended to test and compare FER models, obtaining a metric as a first approximation to a more general and non-specific state-of-the-art as the current one. (3) Perform training and evaluation of CNN-based models on the different datasets presented here. In addition to single-dataset approach, we perform a cross-dataset evaluation considering the merged dataset, unlike related works that use a single dataset different from the training dataset.

The results of single- and cross-dataset experimentation validate our hypotheses. Single-dataset accuracy is higher because the model is evaluated with images from the same dataset, but it is not necessarily a measure of a the model performance on other datasets. It is a specific metric of little or no use in real environments. Cross-dataset accuracy is lower due to the discrepancy between the training and test datasets. In the typical cross-dataset evaluation there is only one test

dataset, so it is not a reliable measure of generalization. In contrast, we evaluate the model on the combined, balanced and unbiased dataset, which provides a more representative metric that can be used as a benchmark for in-the-wild FER. Our approach shows that training on the multiple dataset achieves higher generalization outperforming FER2013 by 13.93%, 24.17% over NHFI and 7.45% over AffectNet. Thus, we identify the most suitable model for the emotion recognition task in real-world applications, which is used for our FER system.

We use a text-image model for the generation of synthetic images and design a proper structure of the prompt, which is the key to obtain a good result. This technique is suitable for dealing with imbalance and bias in facial image datasets. It produces images of high quality and realism, and in less time compared to traditional techniques such as GAN. The identities are not real, which avoids privacy issues. These advantages motivate the creation of a completely artificial dataset of categorical emotions, whose test subset is balanced. Although the single-dataset performance of the artificial dataset is comparable to the real datasets, the cross-dataset performance is very low. This suggests the impossibility of completely replacing the real facial images with the synthetic ones, however, their complementarity has been of great help in this work, given that a small amount of synthetic data is needed. The artificial dataset is a novel product for FER created with AI to assist AI, which promises to reduce costs, time and effort, and improve data quality, in contrast to the classical paradigm of manual dataset collection.

VI. FUTURE WORK

The merged dataset can be extended with more in-the-wild datasets and balanced with the proposed method to train a more general FER model. It is important to combine previously refined datasets and with different properties of size, resolution, color, background, illumination and image format to achieve more variability and allow a model to learn patterns that can be generalized to new data and situations. Although the realism and quality of the synthetic images obtained are very good, the problem of emotions is not yet fully conceivable as a text-image pairing. However, the use of CLIP-based Stable Diffusion offers very promising advantages, such as ease of use, increased speed, automatic and reliable labeling and feature control. These tools are maturing and will be key to understanding the particularities of synthetic images, bridging the gap between the real and artificial domains. We believe that a promising direction of work would be to retrain and tune Stable Diffusion with facial datasets and the proposed prompts, rather than using the model only as an inference tool, which could improve emotion recognition. This would allow the integration of synthetic and real datasets in the training phase and not only in the evaluation phase. Transformer-based architectures are suggested to be tested separately or combined with CNNs. Also, consider facial images with profile postures and combine this work with multimodal emotion recognition systems.

APPENDIX A USE CASE: WEB-BASED FER APPLICATION

We use our generalized model in academic and professional education scenarios, one of the most important real-world applications of FER. Facial expressions are the best evidence of a person's emotional state, which is key in the classroom and office, both physically and virtually. Students' gestures serve as feedback to the teacher to detect engagement or lack of interest, whereas teacher's gestures can be a warning of compliance or non-compliance with educational objectives [61]. The same is true for a company's employees at the time of professional training. Currently, a FER system is a necessity in the growing online or virtual learning, which has also been the cause of many students dropping out, lack of engagement and lower academic quality.

We deployed a Web-based FER system[35] with the main objective of availability, public access and ease of use. The input is a video of a class or meeting that can take place in Google Meet, Microsoft Teams, Zoom or any other similar platform. This input is processed frame by frame by the CNN model obtained from the training of the merged dataset, and as output we can instantly observe the emotion category of each participant. This can be analyzed and associated with the level of attention and concentration of the students or employees, which becomes a support for decision making by the teacher or supervisor. The system operates in post-processing mode, i.e., after the event, but can also be integrated into security cameras for real-time or background monitoring, which will be the subject of further work. For the implementation, we use the popular Flask[36] framework written in Python that allows the development of Web applications.

Figure 16 is a screenshot of our system in action. We select an example of company training in Zoom with people with Asian features, with whom FER systems often have problems recognizing facial expressions, but we show that the results are satisfactory. We can appreciate how the facial region of each participant is framed and assigned the emotion category

[35]https://github.com/cimejia/novel-FER-datasets/tree/main/VideoFERwebApp

[36]https://flask.palletsprojects.com/en/2.3.x/

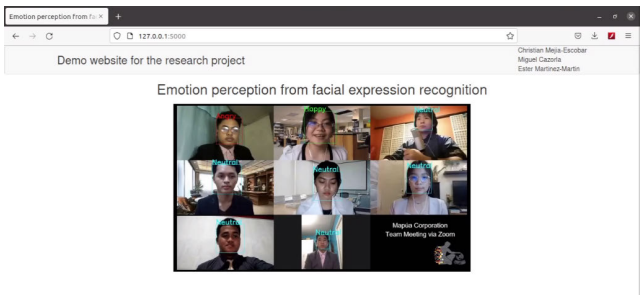
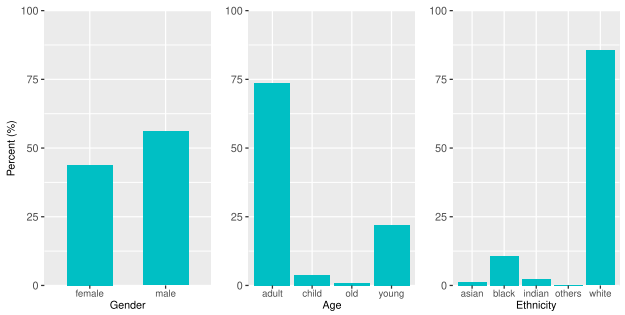
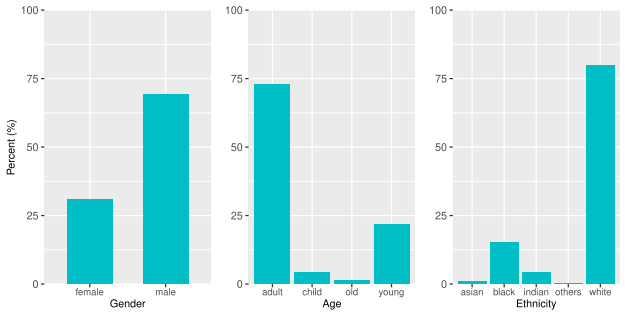


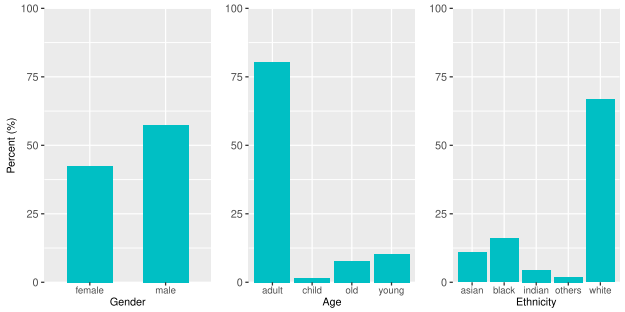
FIGURE 16. Web-based FER system screenshot.



(a) FER2013



(b) NHFI



(c) AffectNet

FIGURE 17. Distribution of facial images according to predictor of gender, age and ethnicity for the datasets: (a) FER2013, (b) NHFI and (c) AffectNet.

label. It is even possible to distinguish by color the type of emotion, where the positive or neutral ones have a light color, whereas the negative ones, such as the angry case, stand out by the red color, which would be immediately perceptible to the person monitoring people's attention and engagement during the class or meeting, as well as a warning to make a decision.

APPENDIX B GENDER, AGE AND ETHNICITY IMBALANCE

We join the prediction files in a spreadsheet[37] and produce the distribution plots shown in Figure 17.

The high imbalance between groups of each variable is evident. Male gender, adult age and white ethnicity predominate in the three datasets. Models trained on these datasets will exhibit a bias favoring these groups. By a wide margin, young age and black ethnicity follow. Finally, there is little presence of children and old people, as well as Asians, Indians and other non-white and non-black ethnicities. These results confirm the well-known lack of representativity of the in-the-wild FER datasets, with overrepresentation of certain groups of people and underrepresentation of others. This is one of the major problems that reduce the accuracy of emotion recognition in different datasets and real environments.

[37] https://github.com/cimejia/novel-FER-datasets/tree/main/FER-evaluator-dataset/AGR-prediction

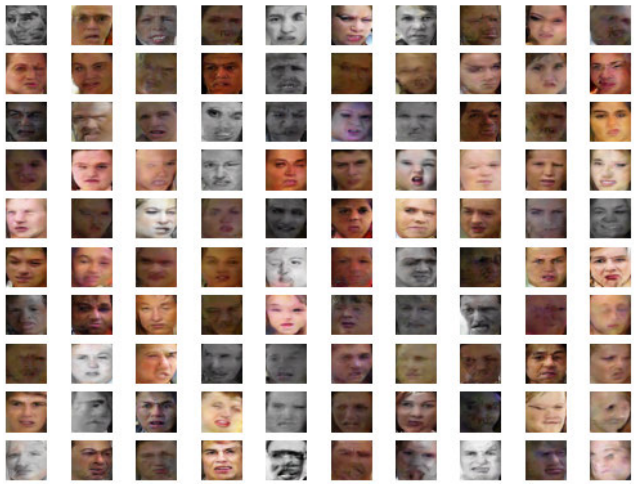


FIGURE 18. Sample of face images for "disgust" generated by the GAN.

FIGURE 19. Face images generated by Stable Diffusion for the angry category.

APPENDIX C SYNTHETIC IMAGES

See Figs. 18 and 19.

APPENDIX D PROMPTING

See Table 11.

APPENDIX E COMPUTATIONAL PLATFORM

The main characteristics of the computational platform used for our experiments are: Intel(R) Core(TM) i9-7920X processor, 2.90GHz, RAM of 64 GB, NVIDIA GeForce RTX2080 GPU with RAM of 12 GB and Linux Ubuntu 18.04.5 LTS. *Python* with *TensorFlow* and *Keras* for CNNs implementation and training. *Sklearn* for metrics and confusion matrix. Libraries such as *OS*, *NumPy* and *Matplotlib* for file system management, numerical arrays and plot display,

TABLE 11. Prompts used for the generation of facial images of emotions with Stable Diffusion.

Emotion	Prompt	Effectiveness
Angry	"A detailed photographic portrait of a perfect face of a <person>, feeling an extreme rage, expressing a very angry face, features well-defined, facing the camera, realistic, 4K, hd"	40%
Disgust	"A detailed photographic portrait of a perfect face of a <person>, feeling angry, with expression of very disgusted, forehead wrinkler, brow lowerer, cheek raiser, narrowed eyes, nose wrinkler, upper lip raiser, chin raiser, lip part, facing the camera, realistic, 4K, hd"	30%
Fear	"A detailed photographic portrait of a perfect face of a <person>, expressing great dread, with facial gestures of fearful, fright, in panic, facing the camera, realistic, 4K, hd"	60%
Happy	"A detailed photo portrait of a perfect whole front face of a <person>, expressing very happiness, facial gestures of happy, smiling, facial features well-defined, facing the camera, background any, ultra realistic, 4K, hd"	90%
Neutral	"A detailed photo portrait of a perfect face of a <person>, expressing neutrality, facial gestures of neutral, facial features well-defined, facing the camera, background any, ultra realistic, 4K, hd"	80%
Sad	"A detailed photo portrait of a perfect whole front face of a <person>, very sad face with tears, expressing extreme frustration, crying, facial features well-defined, facing the camera, background any, ultra realistic, 4K, hd"	90%
Surprise	"A detailed photographic portrait of a perfect face of a <person>, expressive face of surprise with the mouth open, extremely amazed, eyebrows very raised, upper eyelid raised, lips parted, jaw dropped, facial features well-defined, facing the camera, background any, realistic, 4K, hd"	50%

respectively. The *ImageDataGenerator* utility for image processing, dataset splitting and pixel normalization.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] B. Mondal, "Artificial intelligence: State of the art," in *Recent Trends and Advances in Artificial Intelligence and Internet of Things* (Intelligent Systems Reference Library), vol. 172, V. Balas, R. Kumar, and R. Srivastava, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-32644-9_32.

[2] G. Gigerenzer, *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*. Cambridge, MA, USA: MIT Press, 2022.

[3] M. A. Khder, A. Y. Bahar, and S. W. Fujo, "Effect of artificial intelligence in the field of games on humanity," in *Proc. ASU Int. Conf. Emerg. Technol. Sustainability Intell. Syst. (ICETSIS)*, Jun. 2022, pp. 199–204.

[4] H. Hwang and D. Matsumoto, "Functions of emotions," in *Noba Textbook Series: Psychology*, R. Biswas-Diener and E. Diener, Eds. Champaign, IL, USA: DEF Publishers, 2023. [Online]. Available: http://noba.to/w64szjxu

[5] L.-F. Chen, M. Wu, W. Pedrycz, and K. Hirota, *Emotion Recognition and Understanding for Emotional Human-Robot Interaction Systems*, 1st and 2021st ed. Springer, Nov. 2020.

[6] L. Fiorini, F. G. Loizzo, G. D'Onofrio, A. Sorrentino, F. Ciccone, S. Russo, F. Giuliani, D. Sancarlo, and F. Cavallo, "Can i feel you? Recognizing human's emotions during human–robot interaction," in *Social Robotics*. Florence, Italy: Springer, Dec. 2023, pp. 511–521.

[7] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human–robot interaction: Recent advances and future perspectives," *Frontiers Robot. AI*, vol. 7, Dec. 2020, Art. no. s532279. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.532279

[8] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intell. Syst. Appl.*, vol. 17, Feb. 2023, Art. no. 200171.

[9] T. Ghosh, M. H. A. Banna, M. J. A. Nahian, M. S. Kaiser, M. Mahmud, S. Li, and N. Pillay, "A privacy-preserving federated-mobilenet for facial expression detection from images," in *Applied Intelligence and Informatics*. Reggio Calabria, Italy: Springer, Sep. 2023, pp. 277–292.

[10] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," 2016, *arXiv:1608.01041*.

[11] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial emotion recognition with vision transformers," *Appl. Syst. Innov.*, vol. 5, no. 4, p. 80, Aug. 2022.

[12] Y. Jiao, Y. Niu, T. D. Tran, and G. Shi, "2D+3D facial expression recognition via discriminative dynamic range enhancement and multi-scale learning," 2020, *arXiv:2011.08333*.

[13] W. Dias, F. Andaló, R. Padilha, G. Bertocco, W. Almeida, P. Costa, and A. Rocha, "Cross-dataset emotion recognition from facial expressions through convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 82, Jan. 2022, Art. no. 103395. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320321002637

[14] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[15] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion*, vols. 83–84, pp. 19–52, Jul. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522000367

[16] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1609–1618.

[17] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system (FACS). A technique for the measurement of facial action," Consulting, Palo Alto, CA, USA, 1978, p. 22.

[18] E. L. Rosenberg and P. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford, U.K.: Oxford Univ. Press, 2020.

[19] J. Yang, J. Shen, Y. Lin, Y. Hristov, and M. Pantic, "FAN-trans: Online knowledge distillation for facial action unit detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6008–6016.

[20] A. Greco, N. Strisciuglio, M. Vento, and V. Vigilante, "Benchmarking deep networks for facial emotion recognition in the wild," *Multimedia Tools Appl.*, vol. 82, pp. 11189—11220, Mar. 2023.

[21] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.

[22] 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. 2019.

[23] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou, "Deep neural network augmentation: Generating faces for affect analysis," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1455–1484, May 2020, doi: 10.1007/s11263-020-01304-3.

[24] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, p. 268, May 2022.

[25] H. Meng, F. Yuan, Y. Tian, and T. Yan, "Cross-datasets facial expression recognition via distance metric learning and teacher–student model," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5621–5643, Feb. 2022.

[26] C. Mejia-Escobar, M. Cazorla, and E. Martinez-Martin, "Towards a better performance in facial expression recognition: A data-centric approach," in *Computational Intelligence and Neuroscience. Advances in the Application of Human Activity Recognition*. Hindawi, 2023.
[27] S. Ramis, J. M. Buades, F. J. Perales, and C. Manresa-Yee, "A novel approach to cross dataset studies in facial expression recognition," *Multimedia Tools Appl.*, vol. 81, pp. 39507–39544, Apr. 2022.
[28] R. Gozalo-Brizuela and E. C. Garrido-Merchan, "ChatGPT is not all you need. A state of the art review of large generative AI models," 2023, *arXiv:2301.04655*.
[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021, *arXiv:2112.10752*.
[30] S. Banerjee, J. S. Bernhard, W. J. Scheirer, K. W. Bowyer, and P. J. Flynn, "SREFI: Synthesis of realistic example face images," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 37–45.
[31] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," 2019, *arXiv:1912.04958*.
[32] R. A. Zafra, L. A. Abdullah, R. Alaraj, R. Albezreh, T. Barhoum, and K. A. Jallad, "An experimental study in real-time facial emotion recognition on 3RL dataset," *J. Current Trends Comput. Sci. Res.*, 2022, doi: 10.33140/jctcsr.
[33] J. H. Kim and D. S. Han, "Data augmentation & merging dataset for facial emotion recognition," in *Proc. Symp. 1st Korea Artif. Intell. Conf.*, Jeju, South Korea, 2020, pp. 12–16.
[34] H. Gao and K. Ogawara, "Face alignment by learning from small real datasets and large synthetic datasets," in *Proc. Asia Conf. Algorithms, Comput. Mach. Learn. (CACML)*, Mar. 2022, pp. 397–402.
[35] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "SFace: Privacy-friendly and accurate face recognition using synthetic data," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–11.
[36] D. Nikolova, I. Vladimirov, and Z. Terneva, "Artificial humans: An overview of photorealistic synthetic datasets and possible applications," in *Proc. 57th Int. Scientific Conf. Inf., Commun. Energy Syst. Technol. (ICEST)*, Jun. 2022, pp. 1–4.
[37] X. Jin, W. Sun, and Z. Jin, "A discriminative deep association learning for facial expression recognition," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 4, pp. 779–793, Apr. 2020.
[38] V. Vonikakis, N. Y. R. Dexter, and S. Winkler, "Morphset: Augmenting categorical emotion datasets with dimensional affect labels using face morphing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2713–2717.
[39] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, M. Johnson, V. Estellers, T. J. Cashman, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone," Work Conducted Mixed Reality AI Lab-Cambridge, Microsoft, Tech. Rep., 2021.
[40] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "DigiFace-1M: 1 million digital face images for face recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Oct. 2023, pp. 3515–3524.
[41] A. Kumar, L. Bi, J. Kim, and D. D. Feng, "Chapter five—Machine learning in medical imaging," in *Biomedical Information Technology* (Biomedical Engineering), 2nd ed. D. D. Feng, Ed. New York, NY, USA: Academic Press, 2020, pp. 167–196. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128160343000055
[42] L. Colbois, T. D. F. Pereira, and S. Marcel, "On the use of automatically generated synthetic image datasets for benchmarking face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.
[43] B. Bozorgtabar, M. S. Rad, H. K. Ekenel, and J.-P. Thiran, "Using photorealistic face synthesis and domain adaptation to improve facial expression analysis," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Oct. 2019, pp. 1–8.
[44] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3D imitative-contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5154–5163.
[45] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "SynFace: Face recognition with synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2021, pp. 10860–10870, doi: 10.1109/ICCV48922.2021.01070.
[46] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.
[47] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Advances in Neural Information Processing Systems*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds. Curran, 2021, pp. 852–863. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf
[48] BwandoWando. (2022). *Face Dataset of People That Don'T Exist*. [Online]. Available: https://www.kaggle.com/dsv/4152463
[49] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.
[50] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf
[51] D. Beniaguev. (2022). *Synthetic Faces High Quality (SFHQ) Dataset*. [Online]. Available: https://github.com/SelfishGene/SFHQ-dataset
[52] BwandoWando. (2022). *Face Dataset Using Stable Diffusion V.1.4*. [Online]. Available: https://www.kaggle.com/datasets/bwandowando/faces-dataset-using-stable-diffusion-v14
[53] G. Sunitha, K. Geetha, S. Neelakandan, A. K. S. Pundir, S. Hemalatha, and V. Kumar, "Intelligent deep learning based ethnicity recognition and classification using facial images," *Image Vis. Comput.*, vol. 121, May 2022, Art. no. 104404. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885622000336
[54] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais, and T. Yasir, "Age and gender prediction using deep convolutional neural networks," in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Dubai, Nov. 2019, pp. 1–6.
[55] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.
[56] J. Brownlee, *Generative Adversarial Networks With Python: Deep Learning Generative Models for Image Synthesis and Image Translation*. Vermont, VIC, Australia: Machine Learning Mastery, 2019.
[57] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5b: An open large-scale dataset for training next generation image-text models," in *Proc. 36th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2022, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=M3Y74vmsMcY
[58] A. Bhalla. (2020). *Facial Expression Recognition*. [Online]. Available: https://www.kaggle.com/code/bhallaakshit/facial-expression-recognition/
[59] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[60] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[61] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct. 2022.

CHRISTIAN MEJIA-ESCOBAR received the B.S. degree in computer and information systems engineering from National Polytechnic University (EPN), Quito, Ecuador, in 1999, the M.Sc. degree in computer science from the Center for Research and Advanced Studies, National Polytechnic Institute (CINVESTAV), Mexico, in 2007, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2023. Since 2013, he has been a Professor with the Central University of Ecuador, Quito, where has been a full-time Professor in software and statistics, since 2016. His research interests include machine learning and deep learning, especially in computer vision tasks.

MIGUEL CAZORLA (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the University of Alicante, in 2000. In 1995, he joined the University of Alicante as a Professor. He was a Computer Engineer with the University of Alicante, in 1995. Since 2017, he has been a Full Professor with the University of Alicante. He has completed several stays at foreign institutions (Carnegie Mellon University, the University of Sydney, and The University of Edinburgh). He has published more than 50 articles indexed in JCR (with more than 20 in Q1) and more than 100 publications in national and international conferences. His research interests include computational vision to solve robotic tasks, social robotics to help dependents, the processing of 3D data, and deep learning to different areas, such as medical image, object recognition, depth estimation, and the identification of traffic objects. He is a member of different program committees of national and international conferences.

ESTER MARTINEZ-MARTIN (Senior Member, IEEE) received the B.S. degree in computer science engineering, the master's degree in secondary education, vocational training and language teaching, the Ph.D. degree in engineering (robotics), and the master's degree in mobile and video games programming from Jaume-I University, in 2004, 2011, 2011, and 2017, respectively. She is currently an Associate Professor with the University of Alicante and the Assistant Director of the University Institute for Computing Research. She has published 14 JCR-indexed articles, five articles in journals (Scopus-indexed), a research book (Springer), several book chapters, three research books (as a co-editor), and more than 25 articles in conferences, both national and international. She has done four research stays with Università degli Studi di Genova, Sungkyunkwan University, Universidade do Minho, and Technische Universitat (TU) Wien. She is a member of AERFAI. Her research interest includes the use of vision in robotic tasks, such as object detection and action recognition. She has been a member of the program committee and organization committee in several national and international conferences.