A NEW NEURAL NETWORK ARCHITECTURE FOR FACIAL EXPRESSION RECOGNITION

Emanuele Raimondo
Department of Computer Engineering
Politecnico di Torino
C.so Duca degli Abruzzi 24,
Torino 10129, Italy

Elena Carlotta Olivetti
Department of Management and Production Engineering
Politecnico di Torino
C.so Duca degli Abruzzi 24,
Torino 10129, Italy

Leonardo Tanzi
Methinks AI
Pça. de Pau Vila, Ciutat Vella,
Barcelona 08003, Spain

Enrico Vezzeti
Department of Management and Production Engineering
Politecnico di Torino
C.so Duca degli Abruzzi 24,
Torino 10129, Italy

Federica Marcolin
Department of Management and Production Engineering
Politecnico di Torino
C.so Duca degli Abruzzi 24,
Torino 10129, Italy

***Abstract*— Facial Emotion Recognition (FER) is a crucial and complex task in computer vision with applications across numerous fields such as human-computer interaction, psychology, healthcare, and marketing. FER aims to accurately detect and interpret a person's emotional state based on facial expressions. However, significant challenges arise due to the inherent variability of expressions across different individuals, cultural backgrounds, and contexts. While recent advances in deep learning, particularly in Convolutional Neural Networks (CNNs), have significantly improved the accuracy of FER systems, these models still struggle to generalize well across diverse datasets, especially when facial emotions are subtle or mixed.**

This paper introduces a novel architecture for 3D multimodal facial expression recognition, which leverages a combination of CNNs for feature extraction, a spatial attention network applied in cross modality and a transformer encoder for contextual representation. By integrating these modalities, the proposed system effectively captures spatial features from facial expressions, addressing challenges related to subtle emotions. To validate the proposed approach, extensive experiments were conducted on a widely recognized dataset (BU3DFE), and a novel dataset (CalD3rMenD3s). A final experiment is conducted on a merged dataset which is currently the largest static dataset for 3D facial expression recognition in the literature.

The proposed architecture demonstrates the potential of multimodal architectures in advancing the field of FER showing competitive performance compared to existing methods.

I. INTRODUCTION

Facial expression recognition (FER) is the automated process of identifying and categorizing facial expressions in images or videos. This task involves interpreting emotions such as happiness, sadness, anger, surprise, disgust, and fear based on facial features. FER has become essential in human-computer interaction, enabling systems to understand and respond to human emotions, enhancing user experience across diverse applications, including virtual assistants, gaming, and social robotics.

While facial expressions are a primary modality for conveying emotions, other factors such as body movements, voice, and physiological signals also play a role in emotion recognition. Despite these, FER remains one of the most widely researched and applied aspects of non-verbal communication in affective computing due to its relative ease of acquisition and analysis.

The nature of FER datasets is pivotal to the advancement of this field. Datasets differ based on annotation types and data acquisition methods. Early works, such as Ekman's categorical model [13], proposed six basic emotions (anger, surprise, happiness, fear, sadness, and disgust), asserting that these emotions are universally recognized across cultures. However, recent psychological research suggests that emotional expressions can vary in intensity and frequency across different cultures, challenging the universality of categorical models [22]. To address this, alternative annotation models, such as the **Facial Action Coding System** (**FACS**), have been introduced. FACS, in particular, breaks down facial expressions into muscle movements called Action Units (AUs), offering a more granular representation of expressions and their physical manifestations. This approach is valuable for cross-cultural analysis, as it focuses on the mechanical aspects of facial movement rather than subjective emotional experience[14].

Datasets also differ in terms of the type of data collected—whether **static** or **dynamic**, **posed** or **spontaneous**, **2D** or **3D**, and acquired in controlled **laboratory settings** or **in-the-wild** environments. Static datasets consist of single images capturing spatial information, while dynamic datasets

capture temporal evolution, allowing for the extraction of both spatial and temporal features. The latter is particularly useful for understanding the subtleties of expressions over time. Posed datasets, where subjects are instructed to display specific expressions, are easier to collect but may not reflect real-world scenarios. Conversely, spontaneous datasets contain naturally occurring expressions, which are more representative of real-life emotions but are more expensive and challenging to collect. Figure 1 provides an example of some common FER datasets, highlighting their characteristics.

The selection of a dataset has a significant impact on the performance and generalizability of FER models. Moreover, the integration of multimodal data, including audio and physiological signals, enhances model robustness, especially in dynamic datasets, by providing complementary information for emotion recognition. The increasing use of 3D representations, such as depth maps and point clouds, also helps mitigate challenges like illumination variation and occlusion, though they come at a higher cost compared to 2D representations.

As FER continues to evolve, the demand for datasets that better reflect the diversity and complexity of human emotion is growing. The global emotion recognition market, valued at 21.7 billion in 2021, is projected to reach 136.2 billion by 2031, with a compound annual growth rate (CAGR) of 20.5% [49], underscoring the expanding interest and applications of FER technology across industries.

II. Related Works and challenges

Savchenko et al. [38][39] developed multi-task, CNN based models using architectures like MobileNet [19] and EfficientNet [48] and RexNet [15]. They demonstrated such lightweighted and fast backbones are enough to achieve state of the art performances in a multitude of tasks (face verification, face recognition and FER) providing a strong baseline for many applications. Such multitask networks are usually made of multiple streams (one or more per sub-task) and they compute a final loss by combining the losses of the individual subtask (usually with a sum). In particular, in [39], they pre-train different lightweight CNNs over VGGFace2 [7] (for face recognition) and, in a second experiment, over ImageNet [37] datasets (for object recognition). The pretraining is useful to let the network produce features suitable to discriminate subjects from another. The backbone is appended with 4 stream networks (one per subtask) and finetuned over AffectNet [33]. These 4 stream networks are FC layers for gender, age, ethnicity recognition subtasks, while the one for FER subtask also containes some additional convolutional layer. This is because FER is a different task from attribute recognition, so a common approach is to extract earlier features from the backbone CNN and fine-tune the convolutional layers for the downstream task (FER). They show that pretrained backbone over VGGFace2 leads to better accuracies (˜+4%) in FER task over AffectNet with respect to the same backbone pretrained over ImageNet. This is because VGGface2 is a dataset specifically designed for face analysis tasks, so the backbone CNN will learn to extract more relevant features for such tasks.

Another CNN based work is DDAMFN by Zhang et al. [60]. Similarly to Savchenko et al. they decide to use lightweight CNN backbone as a feature extractor (MobileFaceNet [8]) enhanced for FER task. The common idea raises from the fact that using deeper architectures on small FER tasks could lead to overfitting.
The first enhancement consists in applying multiple-size kernels that permit to capture more diverse features from input images, similarly to the Inception Block from GoogleNet [47]. The second enhancement is the Coordinate Block Attention Module (CBAM)[55] which is used to capture the long-range spatial dependencies between different regions of the feature maps. The CAM is composed of two branches to capture the vertical dependencies and the horizontal dependencies respectively. The two branches are then combined to produce the final attention map which is summed to initial feature map.

Ning et al. in 2024 [35] produced FMAE wich is a Masked Autoencoder (MAE) from [17] trained over Face9M dataset. Face9M is ad hoc created dataset containing 9M images from the unification of common facial datasets used for face analysis tasks (CelebA [28], FFHQ [2], VGGFace2 [7], CASIA-WebFace [57], UMDFaces [3], MegaFace [32], LAION-5B[42], EDFace-Celeb-1M [59]). As with any autoencoder, MAE is composed of an encoder and a decoder. The encoder maps input image into a latent feature of smaller dimension; while the decoder reconstructs the input from the latent feature. Intuitevely, pre-training the encoder in this self-supervised setting allows it to learn features that are relevant for reconstruction, embedding meaningful information for the classification task. After pre-training, the decoder is discarded, and the encoder is used as a feature extractor.

In MAE, the encoder is encouraged to learn useful features by applying random patches to the input image. By omitting some parts of the input image, the encoder is forced to learn feature embeddings invariant to the masked regions. This is particularly useful in computer vision in general, because, differently from language data, image data are not so information-dense and contain a lot of redundancy [17]. Note that this masking idea acts as a regularization (applied at pre-training time) and is very common in FER world to train models more robust to occlusions.[35] Therefore, FMAE [35] pre-trains a Large ViT [12] over Face9M dataset with MAE approach and then use it for FER (and landmark detection) task.Note that FMAE is the best performing model on AffectNet [33] and RAF-DB [25] only using this pre-training and without any complex network architecture or particular attention mechanism. Of course this comes at the high cost of pre-training a large model over a huge dataset.

(a) (b) (c)

Fig. 1: FER datasets examples - (a) CK+[29] sample (grayscale, static, posed, lab acquired), (b) AffectNet[33] sample (RGB, static, spontaneous), (c) 4DFAB[9] sample (3D mesh, dynamic, spontaneous, lab)

Generally speaking, using a CNN backbone for feature extraction and a subsequent attention module to learn relationships from different components in the extracted feature map is a very common strategy in FER. Recently many of these attention modules are becoming transformer based, because they show better ability in evaluating the global context information of the image than CNNs which are limited by the receptive field. This technique, efficiently overcomes the CNNs limitations in capturing long-range dependencies, but comes at the cost of longer training times and larger datasets since learning complex relationships in sequence-like data with transformers is expensive.

For example, in 2023, Mao et al., proposed POSTERV2 [31], a transformer-based model that achieves state of the art performances on RAF-DB and AffectNet. The model is composed of two backbones, the first one extracting features from RGB images (a pre-trained IR-50 [11] that is fine-tuned) and second one that extracts features from landmark information (a pre-trained MobileFaceNet [8] used frozen (non fine-tuned)). The two features are then fused with a transformer used in cross attention setup, which is a transformer that formulates queries using the input from the other modality (e.g., $Q_{RGB} = X_{landmark} \cdot W_{Q_{RGB}}$). This is done 3 times in sequence to extract low, mid and high level features; which are concatenated and fed into a final lightweight ViT [12] (only 2 layers) to learn relationships in the feature embedding and perform classification.

A very similar approach is used by Her et al. in 2024 [18]. They use the same basic architecture with the two stream approach (pre-trained IR-50 for RGB and pre-trained MobileFaceNe for landmarks) with cross attention modules and the final ViT; but they address the annotation ambiguity problem by applying Class Batch Normalization (CBN). According to them, even few ”noisy” (badly annotated) samples can significantly degrade FER performance [18] leading to overfitting, so CBN is used. CBN is a variant of Batch Normalization that normalizes the features of each class separately, rather than the entire batch. In this way they assure that the final features are more class-specific, mitigating the noise from some sample in the batch that may be badly annotated, but also, addressing the high intra-class variance problem.

Another similar model is ARBEX [52] from Wasi et al. that use again the dual stream network, cross attention modules and final ViT.

Xue at al. in TransFER[56] point out the fact that landmark based approaches, like POSTERV2, lack of robustness in wild datasets where landmarks may be not visible because of occlusions, insufficient light or even head pose (for example if the subjects is near profile pose). Another problem is the scaling of the image, as it could be too small to precisely detect landmarks. Finally, the subject's characteristics, such as age and ethnicity, could also be obstacles to consistent landmark detection. Large variations in wrinkles, fat distribution, and muscle tone can make it more difficult to identify key points consistently across different age groups.
TransFER uses a stem CNN (pretrained IR-50) to extract feature from the image. This base feature is then processed by many spatial attention networks which apply 2 1×1 convolutional layers reducing the number of channels to 1. These masks are then Maxpooled, multiplied by original feature map and fed to a transformer encoder to learn relationships between them.
The innovation is in addressing overfitting through a dropout-like strategy. More in depth, one random mask is dropped with probability p1 and also one random head of the transformer is dropped with probability p2, at each forward pass. This strategy is used to prevent the model from overfitting to specific regions of the face and to improve generalization ability.

Huang et al. uses two attention mechanisms in FERVT [21]. The first is a grid-wise attention mechanism that allows the model to extract features focusing on different areas of the face image. The second mechanism is a vision transfomer fed with these low-level features which captures long-range dependencies .

As another example of transformers on top of CNNs, Ma et al. produced VTFF (Visual Transformers with Feature Fusion) [30] which proposes a fusion strategy for RGB and

LBP features (extracted from two backbone ResNet18 [16]) in a global-local attention mechanism that helps the model to be more robust to occlusion and pose variations. More in depth, the features extracted from the two backbones are initially multiplied by two learnable matrixes and then summed. This is done to let the network learn, initially, how much to account for each modality. The overall feature is then processed through a global attention stream (that reduces spatial dimension to 1×1 by Average pooling) and through a local attention stream (which instead retains spatial information). Finally, the attended global and local features are summed and later processed through a large transformer to learn inner relationships between feature components and perform classification. Note that, VTFF actually reaches lower accuracies than other state of the art methods, but the global-local attention express its power when, in original paper, they manually add occlusions to the images demostrating that VTFF is more robust to occlusions than other models.

Li et al. introduced FER-former [26] which is a transformer based model for FER in the wild. Again they try to combine features provided by CNNs with transformers in an hybrid model. The innovation in this work is that, inspired by OpenAI CLIP model [36], they address annotation ambiguity by using soft labels instead of the hard labels which are believed to be noisy. This is particularly true for FER datasets, where annotation ambiguity is usually addressed with a majority voting system, which is very expensive (many expert annotators must be involved).
FER-former comprises a stem feature extractor (they try both ViT and IR-50). These features are then linearly projected and fed to a transformer encoder. In a parallel stream, the soft labels are fed into the pretrained Text encoder from CLIP (which is a transformer trained on image-text pairs from the internet) to extract context aware features. "Soft labels" refers to labels that embed the original one-hot label but are also able to give context information to the text encoder (e.g., formulations like "This is a face image of {expression} " or passive sentences like "a/an {expression} expression is shown in the image").
Later the cosine similarity between the image feature I and the text feature T is computed and used as loss. In this way, the image feature extractor is trained to produce image features that are close to the text features produced from the text encoder, thereby easing the issue of annotation ambiguity.
At test time, the text encoder is discarded and the image feature extractor is used as a standard FER model.

Multimodal emotion recognition leverages information from multiple modalities to enhance the accuracy and robustness of emotion recognition systems. By combining information from different sources, multimodal systems can capture a more comprehensive view of human emotions, leading to more accurate and reliable emotion recognition. The most common modalities used in multimodal emotion recognition are RGB images and 3D facial scans. In particular, 3D facial scans have gained popularity due to their ability to capture detailed facial geometry and expression dynamics, making them well-suited for capturing subtle emotional cues that may be missed in 2D images. 3D representations offer a more robust description that is invariant to environmental variations irrelevant to the FER task (such as illumination or occlusion variations), but they come at higher collection cost.
The most common 3D representation are **depth maps** which are 2D images that encode the depth information of the face. Other 3D representations like **meshes** and **point clouds** are not very popular because of the high computational cost associated with their processing due to their unordered structure. Instead, depth maps are commonly used as they provide a good trade-off between accuracy and computational cost since they can be easily processed by 2D convolutional neural networks. For example, [46], [61] extract depth maps from 3D meshes and use them as input to convolutional neural networks. When it comes to multimodal networks, the most common approach is to use a **feature concatenation** (or sum) strategies where the features extracted from each modality are concatenated (or summed) and fed to a classifier. However, this approach has some limitations, such as the inability to capture complex interactions between modalities and the risk of overfitting due to the high dimensionality of the concatenated features. To address these challenges, some works have explored strategies where features from different modalities are combined at different stages to produce a final embedding containing information from all modalities at early, mid and late stages (e.g., [31], [60]) allowing for more effective integration of information across modalities. However this approach produces very high dimensional features that are difficult to furtherly process for classification. Another approach, which will be of inspiration for this work, is to use **cross-modal attention mechanisms** (e.g., [30]), which enable the model to focus on the most relevant information from each modality, improving the overall performance of the system. These approaches are particularly effective in capturing complex interactions between modalities and use strengh of the information from each modality to improve the performance of the system.

Transformer based models are much more rare for 2D+3D FER because, as already mentioned, transformers generally need more data and there are no large-scale 2D+3D FER datasets due to the high collection costs. This is why most of the works in this field are based on CNNs. To overcome this, Li et al. in MFEVIT[24] in 2021, tried to use a lightweight vision transformers. The same authors later abandoned transformers (probably because prone to overfitting considering the very small sized datasets in 2D+3D FER) in AFNET [46], CMANET [61] and FFNET [45]. They opted instead for ad-hoc developed cross-modality attention fusion networks which are less

”data-hungry” and require less training than transformer.
A common idea used in these works is the application of manually forged masks. Masks are images used as prior knowledge to enhance the CNNs feature extraction. Basically, the network will learn to extract features by focusing on the salient regions indicated by the masks[46] (e.g., eyes, nose, and mouth). The drawback is that, these attention masks can only be effective if all images in the dataset are aligned and in the same format, making it unfeasible for real-world collected dataset or spontaneous dataset where the head rotation may change. In the ablation studies, they get a very small decrease of performance when they drop the depth modality because they use only BU3DFE [58] and Bosphorus [40] as benchmark datasets, which are lab acquired datasets not containing much variation in occlusions, pose and illumination, therefore it is reasonable that the network learns to give much more weight to the RGB modality which is more discriminant in these conditions.

A different approach to solve 2D+3D FER is proposed by Ni et al. in CMFN [34]. Exactly as in above cited works using Bosphorus and BU3DFE datasets, the depth maps are extracted from the 3D meshes, but they convert RGB images into gray scale images and then extract LBP images because LBP features are hand-crafted to be representative of texture information and less affected by illumination changes than RGB. So RGB images are grayscaled to avoid redundant information and let the network focus on the LBP stream.
They extract features from the three modalities (RGB, depth maps and LBP) with specifically developed lightweight CNNs and feed them into a cross-modal fusion network that combines them through attention mechanism similar to the AFNET, CMANET and FFNET.

These attention based fusion networks used in 2D+3D FER perform better than concatenation or sum fusion because, each modality may inherently emphasize different aspects of the facial expression and it is important to let the network learn how much to account each modality for. As mentioned, 2D images capture surface texture and color variations in the face, while depth maps are better at representing subtle facial geometry. So, it is crucial to consider the complementarity of each modality when combining them, such that the model can compensate for weaknesses in one modality with strengths from another. Moreover, it reduces the risk for one modality to dominate the fused feature overshadowing the valuable information provided by the other modality.

Whether multimodal or ”full RGB”, most of the above cited works address the problem of high intra-class variance and high inter-class similarity trough the use of attention mechanisms (specifically developed or transformer based) and use a final standard classification mechanism with softmax and CE loss. Another approach is to make use of loss functions that enhance the discriminative power of the features. For example, Wen et al. in [53] implement center loss to extract well separated and compact features for a face recognition task over the LFW [20] and YTF [54] datasets. CMFN [34] addresses class unmbalance problem using a focal loss.
Also ARBEX [52] uses complex loss function summing the CE loss, Central Loss and an anchor based loss. Lin et al. in [27] use an orthogonalization loss.

Current state of the art solutions, even if presenting similarities, seem to not converge to a unique general solution for FER, but, rather, to a set of solutions specifically designed for different datasets characteristics. For example embedding prior knowledge by using manually forged masks like in [46] works only for the spefific dataset at hand where all images are aligned. In general, more data leads to better generalization, so, despite all the efforts in building robust and efficient architectures, FER will be solved when enough data will be available. For example, the ViT pretrained in a self-supervised manner over 9 milions images (Face9M) dataset in FMAE [35] is able to reach state of the art performances without any particular architecture.

In conclusion, the most promising direction seems to be the use of transformers, but, as already mentioned, they are very data-hungry and require large datasets to be trained. This is why, in this work multiple datasets are merged to try overcome this limitation.

A. Challenges in FER

FER faces several significant challenges that impact its accuracy and effectiveness. Addressing these challenges is essential to improve FER systems, especially given the growing demand for robust applications in real-world environments.

One of the primary challenges is the lack of sufficiently diverse and exhaustive training data. Current state-of-the-art models for FER make use of transformers (e.g., [31][30][56] and many others) which are very ”data-hungry” and require large, high-quality, and diverse datasets to achieve high levels of generalization. However, most available FER datasets are limited in diversity and size, leading to **overfitting**. Overfitting occurs when models memorize the training data, including noise and outliers, but fail to generalize to new, unseen data. To mitigate this, techniques such as data augmentation and the merging of multiple datasets are employed [35]. However, merging datasets is not straightforward, as they often vary in annotation types, collection methodologies, and image quality. Furthermore, spontaneous datasets frequently exhibit imbalanced classes due to the inability of some subjects to adequately display all emotions. Solutions to this issue include adjusting loss functions to handle imbalance and downsampling overrepresented classes during training (e.g. [61] using Focal Loss).

Another critical issue is **annotation ambiguity**, which arises due to the inherent subjectivity of emotions. Human emotions can be difficult to define with precision, leading to

inconsistent labeling, either due to the ambiguous nature of emotional expressions or human error during the annotation process. This annotation noise can significantly impact model performance, as models trained on inconsistent labels may learn non-generalizable patterns . To address this, some datasets rely on multiple annotators, using majority voting to determine the final label [33] [4]. However, even this approach cannot fully eliminate the subjectivity of emotional interpretation, meaning that FER systems must learn to operate under inherent uncertainty. Some approaches involve class specific batch normalization [56] or the use of soft-labels [18] to handle annotation ambiguity.

Inter-class similarity is another common challenge in FER, where certain facial expressions share similar features across different emotion categories. For example, frowning eyebrows may be associated with multiple emotions such as anger, disgust, or contempt. Similarly, emotions like fear and surprise may both involve wide-open eyes, leading to potential confusion for models attempting to distinguish between these categories.

Intra-class variability presents another significant challenge, particularly in datasets collected in uncontrolled environments. Intra-class variability refers to the wide range of variations within the same emotion class, stemming from differences in how individuals express emotions based on factors such as age, ethnicity, gender, and overall expressiveness. Additionally, variations in head pose, lighting conditions, and partial occlusion of facial features can further complicate emotion recognition in unconstrained scenarios.

Addressing *inter-class similarity* and *Intra-class variability* requires models capable of capturing subtle differences between similar expressions, which often involves carefully designed feature extraction techniques and the use of advanced loss function that allow for production of more discriminant features. For example, the **Island Loss** [6] is a popular loss function for learning embeddings that are more discriminative by pushing samples of the same class closer together and samples of different classes further apart. Also data augmentation and other regularization techniques are used to prevent overfitting to specific variations within the training set. Furthermore, model architectures are designed to be robust to these variations, incorporating mechanisms to capture features invariant to changes in pose, lighting, and occlusion.

In this context, additional modalities such as audio or 3D facial scans can be used to provide complementary information that helps disambiguate facial expressions. For example, 3D facial scans provide a more detailed representation of facial features, which can be used to improve the accuracy of facial expression recognition in cases of bad illumination or occlusion. A growing interest in 3D FER has been driven by the need to improve the accuracy of facial expression recognition in real-world scenarios which are much more affected by these variations. This work focuses on 3D facial expression recognition.

In summary, the key challenges faced by FER include the lack of diverse and large-scale datasets, class imbalance, annotation ambiguity, intra-class variability, inter-class similarity. Overcoming these obstacles requires a combination of data augmentation, normalization, regularization, and carefully designed model architectures capable of distinguishing subtle differences between similar emotional expressions. Addressing these issues is crucial for the development of more accurate and reliable FER systems.

B. Contribution

In the following, this work provides a proposal of a deep architecture that is able to deal with multimodal data representation (RGBD) and to learn from the unbalanced dataset. The architecture is based on the state of the art in FER and is designed to be robust to the challenges faced by FER models. Different loss functions are considered evaluating the model's performance on CalD3r, MenD3s [50] and BU3DFE [58] dataset in different conditions. Finally, CalD3r, MenD3s[50], BU3DFE [58] and Bosphorus [40] datasets are merged to create a larger and more diverse dataset and final evaluation of the model's performance. This is the largest static multimodal FER dataset up to our knowledge.

The main contribution of this work consist in answering to the following questions:

- **How to efficiently combine modalities to improve the performance of a FER model?**
  A novel Cross Attention Module is presented
- **How to deal with intra-class variability and inter-class similarity in a multimodal FER model?** Island Loss is used to learn more discriminative features
- **How to create a large and diverse dataset for multimodal FER? And how does a vision transformer based model perform on it?** Multiple datasets are merged to create a larger and more diverse dataset and final evaluation of the model's performance.

III. Materials and Methods

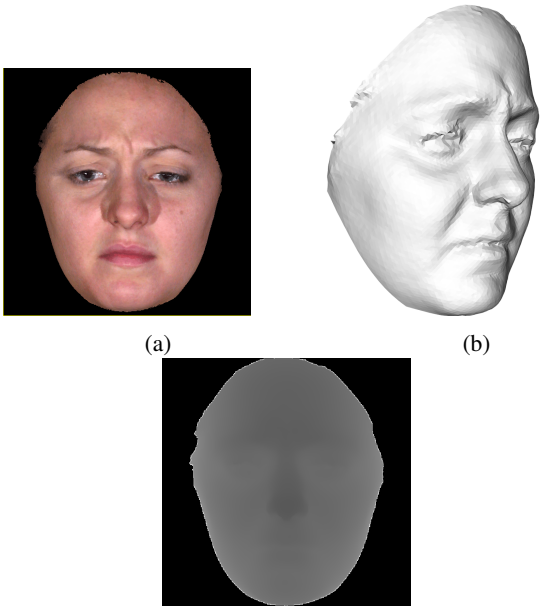
A. Datasets

The main dataset used for validation and comparison is BU3DFE [58]. Later BU3DFE is merged with other two multimodal datasets for FER: CalD3r, MenD3s [50] and Bosphorus [40] and test the model.

BU3DFE The Binghamton University 3D Facial Expression Database (BU3DFE) [58] is a 3D facial expression database that contains 100 subjects with 2500 samples (1300×900) in two modalities: RGB and 3d mesh. The database includes 7 facial expressions: neutral, anger, disgust, fear, happiness, sadness, and surprise. The dataset is posed because subjects are requested to mimic expressions in 4 intensity levels per emotion (except for neutral which only contain 1 intensity level). The resolution is 512×512.

Following common practice, 3D meshes are converted into depth maps by rendering each mesh in 3D space and extracting the depth value. Getting depth maps that

are spatially consistent with corresponding RGB images is challenging because [58] gives no information about the camera device used to capture the RGB image (in particular FOV is not defined). Figure 2 shows an example of reconstruction where we can see the RGB image (a), the corresponding 3D mesh (b) and the depthmap extracted from the mesh (c).



(a) (b) (c)

Fig. 2: (a) RGB, (b) 3D Mesh, (c) extracted Depth Map from BU3DFE [58]

CalD3r and MenD3s In CalD3R there are 104 subjects, 54 woman and 50 men mostly from South Europe with age range 19-35 years old. Instead MenD3S comprises 92 subjects, 46 woman and 46 men from Brazil and age range 18-55. The CalD3r and MenD3s dataset [50] are merged into a unique multimodal dataset containing RGB and depth map modalities. The resulting dataset contains 8617 samples annotated with the basic 7 expressions in 224×224 resolution. The dataset is spontaneous since expressions are elicited by submitting images to subjects from the International Affective Picture System (IAPS) [5] and the Geneva Affective Picture Database (GAPED) [10]. Slight occlusions can be present on some images, such as those caused by hair, beard, and glasses. Different head poses are naturally portrayed. Each sample is described by RGB and Depth map acquired with Intel RealSense SR300 camera.

Bosphorus The Bosphorus dataset [40] is a 3D facial expression dataset that contains 105 subjects with a total of 466 samples annotated in 7 classes. Other images are annotated with Action Units. Following the indications provided in CK+ [29] the most expressive AUs are translated into categorical annotations. Such translation is provided in Table I. Subjects (60 men and 45 women), with age range 25-35, are mostly Caucasian. The RGB image resolution is 1600×1200 and the meshes consist of approximetely 35000 points each. Finally, this is a posed database as the subjects are requested to mimic the expressions.

TABLE I: Bosphorus AUs to Categorical conversion

AUs	Class
Lip Presser (AU24)	Anger
Nose Wrinkler (AU9)	Disgust
Lip Corner Puller (AU12)	Happiness
Lip Corner Depressor (AU15)	Sadness
Outer Brow Raiser (AU2)	Surprise
Inner Brow Raiser (AU1)	Fear

The three datasets distributions are shown in Figure 3.

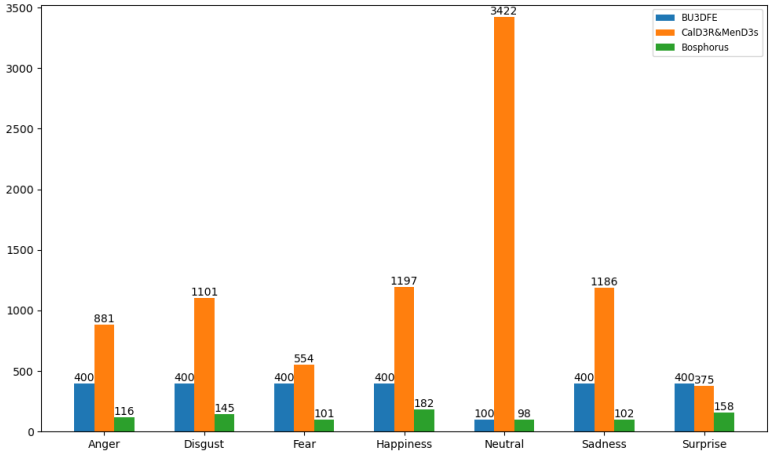


Fig. 3: Classes distribution of the CalD3rMenD3s, BU3DFE and Bosphorus datasets.

B. Preprocessing

In all datasets, faces have been already cropped to remove background. Depth images, consisting of 1 channel, are stacked to reach default network input used for RGB stream. To leverage EfficientNetB2 pre-trained model from [38], the input images are resized to 260×260 with linear interpolation. Pixels values are rescaled into $[0, 1]$ range and z-score normalize as shown in equation 1. Common practice would require to use the mean and standard deviation of the pretraining set (VGGFace2), but that would not guarantee zero centered data for our training set which is very different from VGGFace2, especially for the depth modality. Therefore, at each validation fold, the mean and standard deviation of the training set are computed separately for the two modalities, and used for normalization.

$$\boldsymbol{X_{norm}} = \frac{\boldsymbol{X}/max(\boldsymbol{X})) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{1}$$

Data augmentation will help the model to learn robust patterns to variations in lighting, angle, occlusions, etc..

Furthermore, data augmentation could help reducing inter-class similarity as it provides more diverse examples for each class. On the other hand, introducing overly complex data augmentation could introduce unrealistic noise. Online augmentation is chosen as it provides a continuous stream of diverse variations during training process without requiring additional storage. A limited set of realistic augmentations, are selected: 0°-10° rotations, horizontal flips, color jittering (only for RGB images) which simulates variations in lightning and random erasing which simulates occlusion. Data augmentation is not applied on validation samples as they should represent real world conditions to provide an accurate measure of model performance.

C. Network architecture

Let I_{rgb} and I_{depth} be the RGB and depth images respectively. Resizing and channel stacking for depth images are applied such that $I_{rgb}, I_{depth} \in \mathbb{R}^{H=260 \times W=260 \times C=3}$.

I_{rgb} and I_{depth} are processed in a two stream network in Figure 8, composed of two identical feature extractors (EfficientNetB2) pretrained for face verification on VGGFace2 [39]. The architecture of the EfficientNetB2 backbone is shown in Table II, where $Reps$ is the number of repetitions for a block. Note that kernel size k and stride s, in **Inverted Residual Block** (**IRB**) are referred to the *Depthwise-Conv*, since the *Pointwise-Conv* is always 1×1 with stride 1. Features are extracted after the last IRB leading to $X_{rgb}, X_{depth} \in \mathbb{R}^{H=9 \times W=9 \times C=352}$.

Each Inverted residual block (IRB), shown in Figure 4 contains a **Squeeze-Excite** module which permits to select informative channels and suppress less discriminant ones. The same concept can be applied on the spatial information using **Spatial Attention Mechanisms**. Coordinate Block Attention Module [55] creates the spatial attention map $C \times 1 \times 1$ by concatenationg the MaxPooling and AvgPooling of the feature and then applying convolution and sigmoid activation, as shown in Figure 5(a). Instead, this work uses the LANet [51] approach that uses additional learnable convolution rather than pooling, as shown in Figure 5(b). Similarly to TransFER [56], $S = 4$ spatial Attention Modules are selected and their outputs are MaxPooled, but instead of using drop-out of random attention masks, only batchnorm is used, believing it provides similar regularization effect but without hyperparametrs (no dropout probability).

Moreover, spatial attention is used in a cross modality setup. In particular, the masks M_{rgb} and M_{depth} are computed separately for the two modalities, averaged and multiplied by X_{rgb}. The idea is to select the most informative spatial regions from the RGB and Depth stream and average them such that high activations in the RGB image, which could be due to high local illumination, can be suppressed by low activations in the depth image. Viceversa, high activations in the depth image, which could be caused by occlusions, can be suppressed by low activations in the RGB image. In this way the Depth modality is used to guide the learning of the RGB stream.

The feature X_{out} is passed through a 1×1 convolutional layer and reshaped to create a sequence like input for a small Vision Transformer Encoder (only 4 layers). This setup is the same from POSTERV2 [31]. Finally, two fully connected layers (MLP), followed by a softmax layer are used to classify the output.

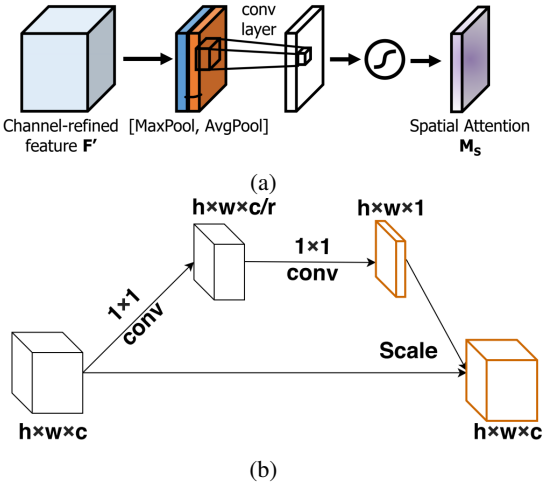


Fig. 5: Spatial Attention mechanisms: (a) CBAM [55], (b) LANet [51]

TABLE II: EfficientNetB2 Architecture

Layer	C_{in}	C_{out}	k	stride	Reps
Conv2D (stem) [a]	3	32	3×3	2	1
Conv2D (stem)[a]	32	16	3×3	2	1
IRB [b]	16	24	3×3	2	1
IRB [b]	24	24	3×3	1	2
IRB [b]	24	48	5×5	2	1
IRB [b]	48	48	5×5	1	2
IRB [b]	48	88	3×3	2	1
IRB [b]	88	88	3×3	1	2
IRB [b]	88	120	5×5	1	1
IRB [b]	120	120	5×5	1	3
IRB [b]	120	208	5×5	2	1
IRB [b]	208	208	5×5	1	4
IRB [b]	208	352	3×3	1	2
IRB [b]	352	352	3×3	1	1
Conv2D[a]	352	1408	1×1	1	1
MaxPool	1408	1408	-	-	1
FC	1408	7	-	-	1

[a] Stem Depthwise Separable block, with batchnorm2D and SiLU

[b] Inverted Residual Block. Residual connection is present only if $C_{in} = C_{out}$

D. Loss Function

The loss layer is a crucial component of the network, as it defines the objective function that the network aims to minimize during training. The choice of the loss function depends on the specific task and the characteristics of the dataset. In FER context, the loss function should take into account the FER's inherent challenges, in particular the high intra-class variance and high inter-class similarity. Furthermore, FER dataset's are usually very unbalanced as

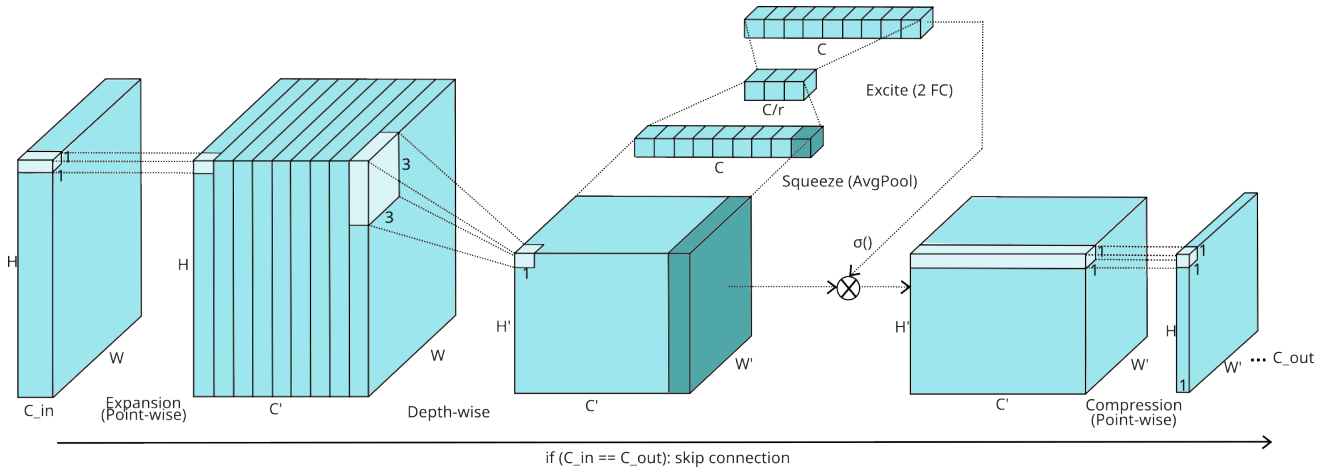


Fig. 4: Inverted Residual Block

some emotions like surprise or contempt are much more infrequent and/or difficult to annotate with respect to others like anger.

In the most common setting, **Cross Entropy Loss** reported in Equation 2, is applied to the output of the softmax layer.

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} w_j \; y_{ij} \; log(p_{ij}) \tag{2}$$

Where N is the number of samples in the training set. Actually, in deep learning N is the number of samples in the mini-batch. K is the number of classes, y_{ij} is the binary true label for sample i and class j ($y_{ij} = 1$ if sample i belongs to class j and $y_{ij} = 0$ otherwise).
Or in vectorized formulation:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w} \cdot \boldsymbol{y}_i \cdot \log(\boldsymbol{p}_i)^T \tag{3}$$

Where $\boldsymbol{y}_i$ is the one-hot encoded label vector for ith sample.

One problem with standard CE loss is that it treats every sample equally, which can be problematic in FER where class imbalance is very common. This can lead to biased models that perform poorly on minority classes. To address this, a weighted version of CE loss can be used, where each class is assigned a weight w_j based on its frequency. This encourages the model to focus more on underrepresented classes, leading to improved performance on imbalanced datasets.

To address high intra-class variability and high inter-class similarity problems, center loss permits the network to produce features that are both separable (featuers from different classes are far apart in the feature space) and discriminant (features from same class are close to each other in the feature space and so encoding the class characteristics). The center loss is reported in Equation 4.

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2 \tag{4}$$

Where $\boldsymbol{c}_{y_i}$ is the class center of the class y_i. Center loss increases the discriminative power of the features by explicitly penalizing the distance between the deep features $\boldsymbol{x}_i$ of each face image and their corresponding class centers in the feature space $\boldsymbol{c}_{y_i}$. Ideally, the class centers should be learnt by computing the mean of the deep features produced at each step for all the samples of the same class in the training set. However, this would be inefficient and impractical. So, class centers are actually updated at each iteration by averaging the deep features of the samples in the mini-batch [53]. This may introduce large perturbations in the learning of the centers (for example, a mini-batch could contain only samples from a single class with a mean very different from the global mean). To avoid this, the learning rate of the centers is controlled by an hyperparameter $\alpha \in [0, 1]$.

The limit of center loss is that it only compresses the clusters individually, but does not push clusters apart. This is why it is used in conjunction with CE loss which forces the fetures of different classes to stay apart.

In 2017, Cai et al. [6] further developed the center loss to produce even more discriminative features and trained a CNN with Island Loss for FER. The combined loss will be therefore the sum of **Island Loss**, center loss and CE loss as presented in Equation 5.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_C + \lambda_2\mathcal{L}_I \tag{5}$$

Where λ_1 and λ_2 are hyperparameters that balance the three loss functions. $\mathcal{L}_I$ is the Island Loss, defined in Equation 6.

$$\mathcal{L}_I = \sum_{\boldsymbol{c}_j}^{K} \sum_{\boldsymbol{c}_k \neq \boldsymbol{c}_j}^{K} \left(\frac{\boldsymbol{c}_j \cdot \boldsymbol{c}_k}{\|\boldsymbol{c}_k\|_2 - \|\boldsymbol{c}_k\|_2} + 1 \right) \tag{6}$$

$\boldsymbol{c}_j$ and $\boldsymbol{c}_k$ are the class centers of class j and k respectively. Intuitively, it minimizes the cosine similarity between the class centers, which encourages the features of different classes to be more separable in the feature space. The +1 term is necessary to make the loss non-negative, since the cosine exists in $[-1, +1]$ range.

Figure 6 shows the features learned by a CNN trained with Cross Entropy loss (a), Center loss (b) and Island loss (c). Note how the center loss aggregates the features of the same expression class towards their centers, thus reducing intra-class variation with respect to the CE loss (a).

The island loss (c) not only compresses the clusters individually but also pushes clusters apart [6].

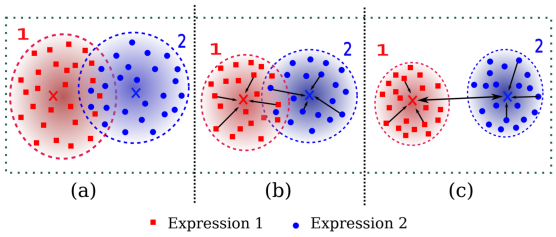


Fig. 6: Features learned by a CNN trained with CE loss (a), Center loss (b), Island loss (c) [6]

Note that other loss functions have been proposed to enhance the discriminative power of the deeply learned features, such as contrastive loss [23], triplet loss [41] which construct the loss function for image pairs and triplets respectively[53]. However, the number of possible pairs and triplets grows quadratically and cubically respectively with the number of samples in the minibatches. A common strategy is to select meaningful pairs and triplets (those that give significant contribute to the training), but these searching algorithm increase training time. In contrast, center loss is computationally efficient and easy to implement, making it a popular choice for enhancing the discriminative power of the features in FER tasks. This comes at the cost of fine-tuning the hyperparameters λ to balance the two loss functions, and α to control the learning rate of the centers.

E. Experimental Settings

Because pretrained backbones are used, the initial weights are biased towards a very different dataset from ours, especially for depth modality. For this reason, common practice suggests to adopt a "warm-up" strategy where the learning rate is initially set to a very low value and then increased to the desired value. This allows the model to slowly adapt to the new dataset without abrupt oscillations caused by large gradients.

A popular learning rate schedule that adopts this strategy is the One-Cycle schedule [44] where the learning rate starts very low, increases up to a maximum value and then decreases following a cosine annealing schedule. Table III shows the learning rate setup for the two backbones and the "fusion network" (comprising the Cross Modality Spatial Attention, the transformer encoder and the MLP) all using One-Cycle schedule. Note that the maximum value is higher for the fusion network because it is trained from scratch, while the backbones are pretrained and only need to be fine-tuned. The learning rate schedule for the fusion network is shown in Figure 7.

TABLE III: Learning rate setup

	Max LR	Min LR	Schedule
Backbones	1^{-4}	10^{-6}	One-cycle
Fusion Network	1^{-3}	10^{-6}	One-cycle

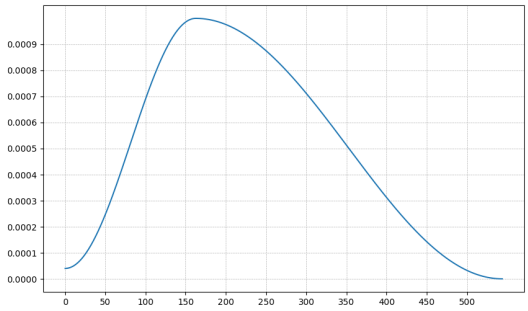


Fig. 7: Onecycle learning rate schedule for the fusion network

AdamW optimizer is used. Island Loss is set up with $\lambda_1 = 1^{-2}$ and $\lambda_2 = 10$, with centers learnt using an SGD optimizer with fixed learning rate $\alpha = 0.5$.

To make the network faster, Automatic Mixed Precision (AMP) is used to make the forward pass in half precision (16bits) while keeping the weights in single precision (32bits). This allows to reduce the memory usage and the computational time.

Batch size is set to 128 with gradient accumulation to make up for memory constraints. Using relatively small batches allows to introduce a regularization effect in the training due to the noise in the gradients computed on smaller batches with respect to larger ones used in other works (e.g. 1024 in [36] or bigger). These values of batch size, learning rate, weight decay, lambdas and S are hyperparameters tuned through a random search strategy (with Optuna framework [1]) to find the best combination of hyperparameters.

The same strategy from AFNET [46], CMANET [61], FFNET [45], MFEVIT [24], CMFN [34], OGFNET [27] (and others) is used to validate the model over BU3DFE. The neutral class, which is under represented, is discarded and only the 2 highest intensity level images are used. Moreover, only 60 random subjects out of 100 are selected. Finally, 10 fold cross validation is conducted over these 60 subjects, for 100 repetitions, selecting different folds at each repetition. The folds accuracies, inside each repetition, are averaged to get repetition accuracy. Later also the 100

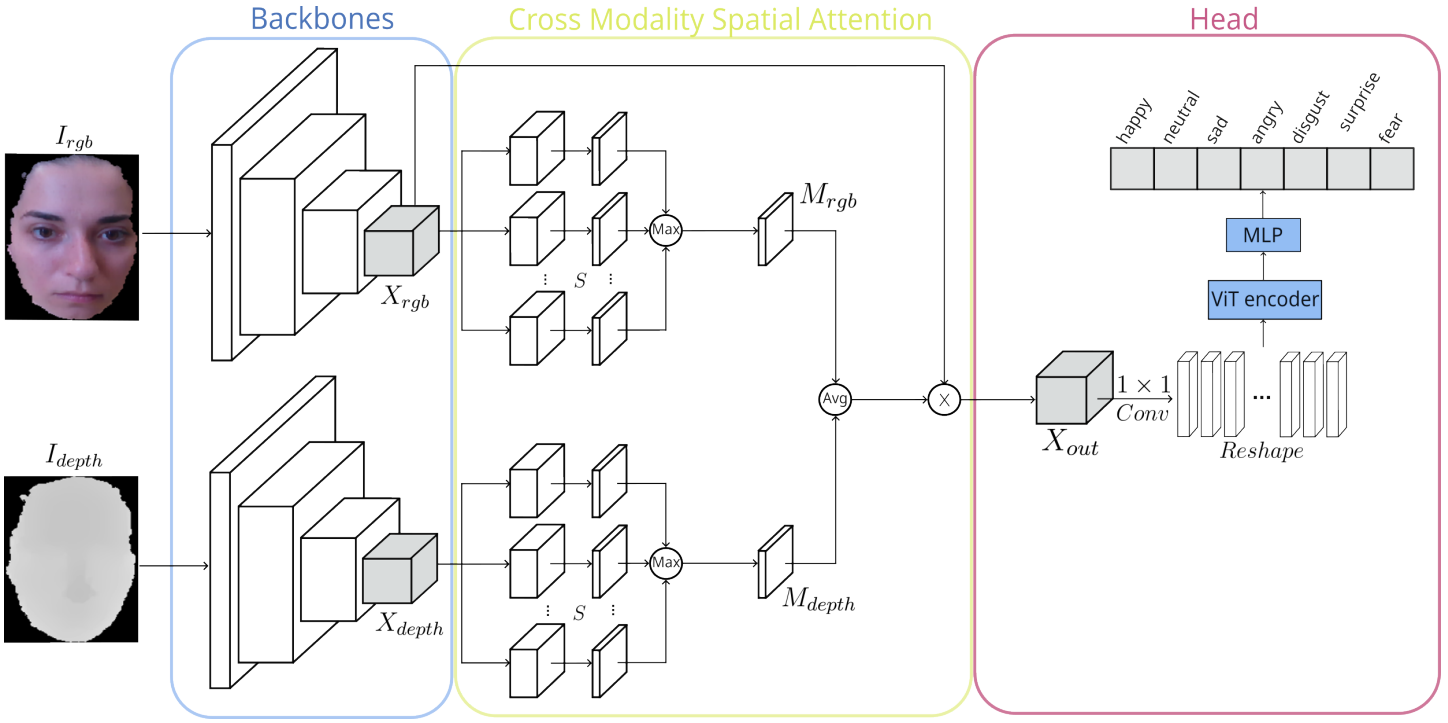


Fig. 8: Model architecture

repetition accuracies are averaged to get the final results. This evaluation protocol is difficult to reproduce because the 60 subjects are selected once at the beginning and it may happen that "easy" subjects are selected. In fact, in order to reproduce the same experiment, the 60 subjects should be the always the same, but the above cited works don't share this information. Moreover, the total amount of images used for validation is only 720 which is very small considering the initial 2500 samples.

Therefore, this work sets a new benchmark for BU3DFE, selecting a 5-fold cross validation over the whole dataset (considering all emotional intensity levels and also the neutral class). Therefore, in this second setup the dataset is larger (2500 samples) and larger batch size (256) can be used. Training is performed for approximately 13 epochs. In every setup, the model is trained for 20 epochs on each fold. Tesla T4 GPU with 12GB memory is used.

IV. Results

Figure 10 and 11 show the evolution of the loss, training and validation accuracy in the two experimental setups. Consistent results are observed across different folds ensuring robustness of this validation strategy. Note that, since gradient accumulation is used, the number of iterations is shown on the x-axis. The translation into epochs is:

$$\text{Epochs} = \frac{\text{Iterations}}{\text{training set size}/\text{Batch size}} \tag{7}$$

For example, in first setup the training set size is 1200×0.6 and batch size is 128 which means that $5,6$ iterations correspond to 1 epoch.

Note that the loss and the validation accuracy are respectively decreasing and increasing consistently over the training, which is a good sign that the model is learning the task without overfitting. Accuracy and confusion matrix are reported in Table IV V and Figure 12. They consist in the averaged values over all the folds. Our model is performing better than current state of the art, with an accuracy of 91,66% with the same experimentation setup of the other cited works.

Note that all the known models validated on BU3DFE are based on a two streams architecture (RGB (or grayscale) and Depth) with different fusion strategies and none of them is using the mesh directly. Looking at the confusion matrices, in the first setup, the worst performing class is Surprise. Note the high overlap between disgust-fear, fear-surprise and disgust-anger; this is because such expressions present similar features like open mouth and wide eyes for Fear-Surprise, or frowned eyebrows Disgust-Anger.

In the second setup, considering also neutral class and lower intensity expressions, the worst performing class is the Neutral , probably because of the low number of neutral samples (only 100 samples) and the high similarity between neutral and lower intensity expressions.

CalD3r and MenD3s is also used for model validation, where the model outperform the Marcolin et al. network [50] from 58,30 to 65,11 (in the 6 classes setup, without surprise) and 62,50 when considering all 7 emotions. The comparison of results between the two datasets confirms that FER over posed datasets, like BU3DFE, is much easier task than over spontaneous dataset, like CalD3rMenD3s. In fact, Figure 9 shows that the features extracted by the model

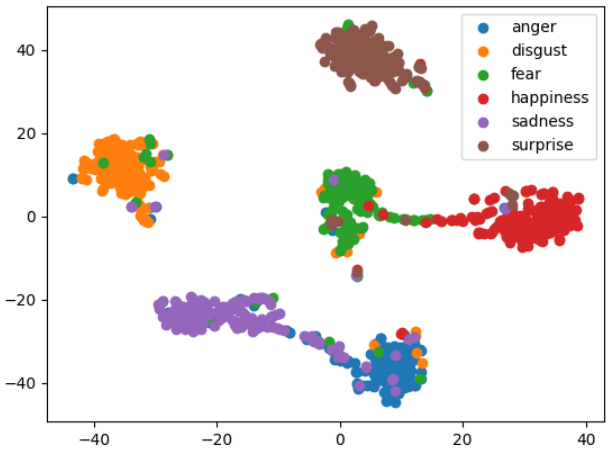


Fig. 9: BU3DFE 6 classes features

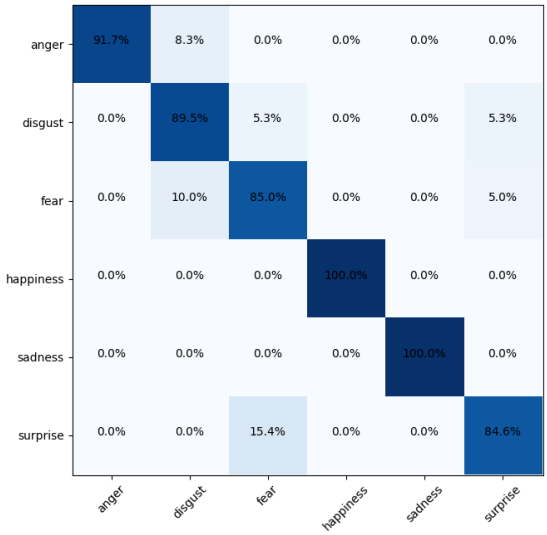
on the BU3DFE dataset are very well separated since the posed expressions are more different from each other with respect to the spontaneous ones in CalD3rMenD3s analysed in Figure 14.

TABLE IV: Results on BU3DFE (w/o Neutral class, only 2 highest intensity levels, 60 subjects. Results averaged on 10 fold validation 100 times)

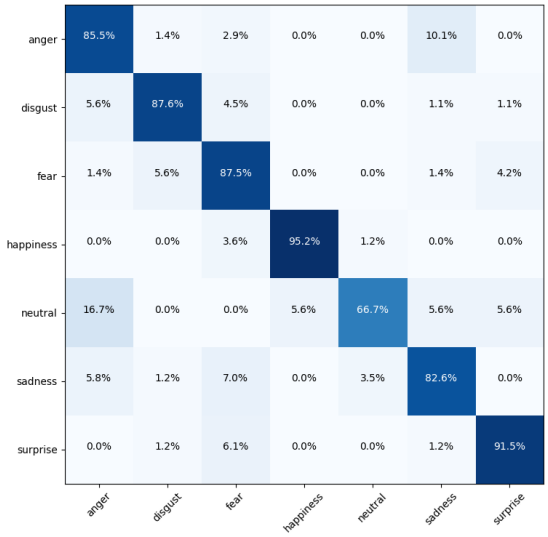
Model	Year	Accuracy (%)	Classes
MFEVIT [24]	2021	90,83	6
CMANET [61]	2022	90,24	6
CMFN [34]	2023	88,91	6
AFNET [46]	2023	90,08	6
This	2024	91,66	6

TABLE V: Our benchmark on BU3DFE (Full dataset. Results averaged on 5 fold validation)

Model	Year	Accuracy (%)	Classes
This	2024	87,6	7



(a) w/o Neutral class, only 2 highest intensity levels, 60 subjects.
Results averaged on 10 fold validation 100 times



(b) Full dataset.
Results averaged on 5 fold validation

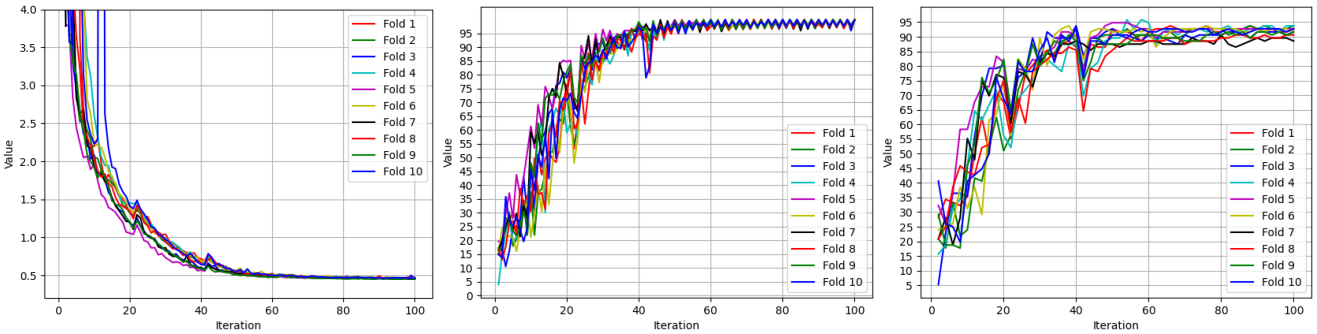
Fig. 12: BU3DFE Confusion Matrices. X-axis are the predicted class, Y-axis is the true class.

A. *Attention analysis*

GRADCAM [43] is used to visualize where the model is focusing on the input image to make the prediction. Gradcam computes the score of the target class at the output of a convolutional layer and then performs backpropagation using that score as a loss value. Figure 13 shows the attention maps for 2 correctly classified validation samples per each class. As expected, note that the most informative regions, where the network poses more attention, are the eyebrows, eyes and mouth.

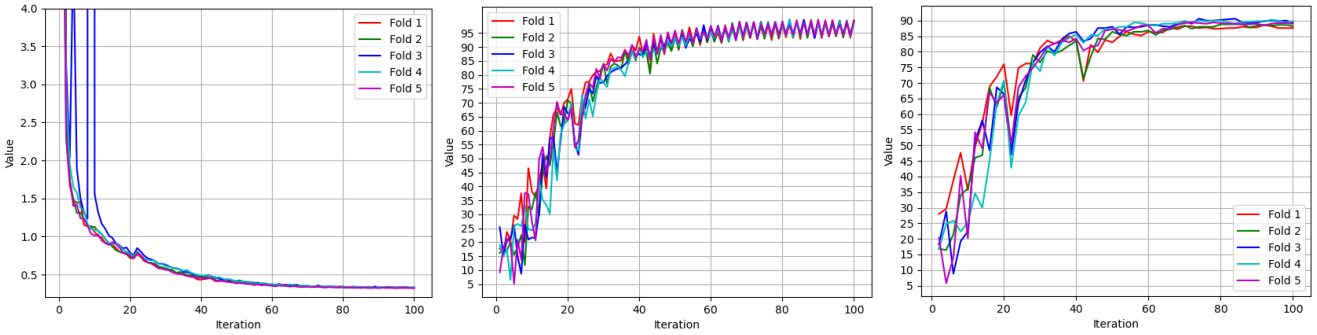
B. *Modality Ablation*

The contribution of depth modality is now evaluated by training the model only using the RGB stream. The results



(a) Training Loss (b) Training Accuracy (c) Validation Accuracy

Fig. 10: Training and Validation Statistics for BU3DFE (w/o Neutral class, only 2 highest intensity levels, 60 subjects). 90-10 validation over 1 repetition as an example.



(a) Training Loss (b) Training Accuracy (c) Validation Accuracy

Fig. 11: Training and Validation Statistics for BU3DFE (Full dataset, 5 fold validation)

shown in Table VI are referred to the same cross validation used for the previous experiments.

TABLE VI: Modality Ablation

Modality	Accuracy (%)
RGB	88,72
RGB+Depth	91,66

Note that the depth modality provides a significant improvement in the accuracy, confirming the importance of the depth information.

C. Modules Ablation

Table VII compares the full model with and without final ViT encoder (which is substituted by a final AvgPooling), and without the Cross Modality Spatial Attention which is replaced with sum fusion between the features from the two modalities:

$$\boldsymbol{p} = Softmax(MLP(AvgPool((\boldsymbol{X_{rgb}}) + (\boldsymbol{X_{depth}})))) \quad (8)$$

TABLE VII: Modules Ablation

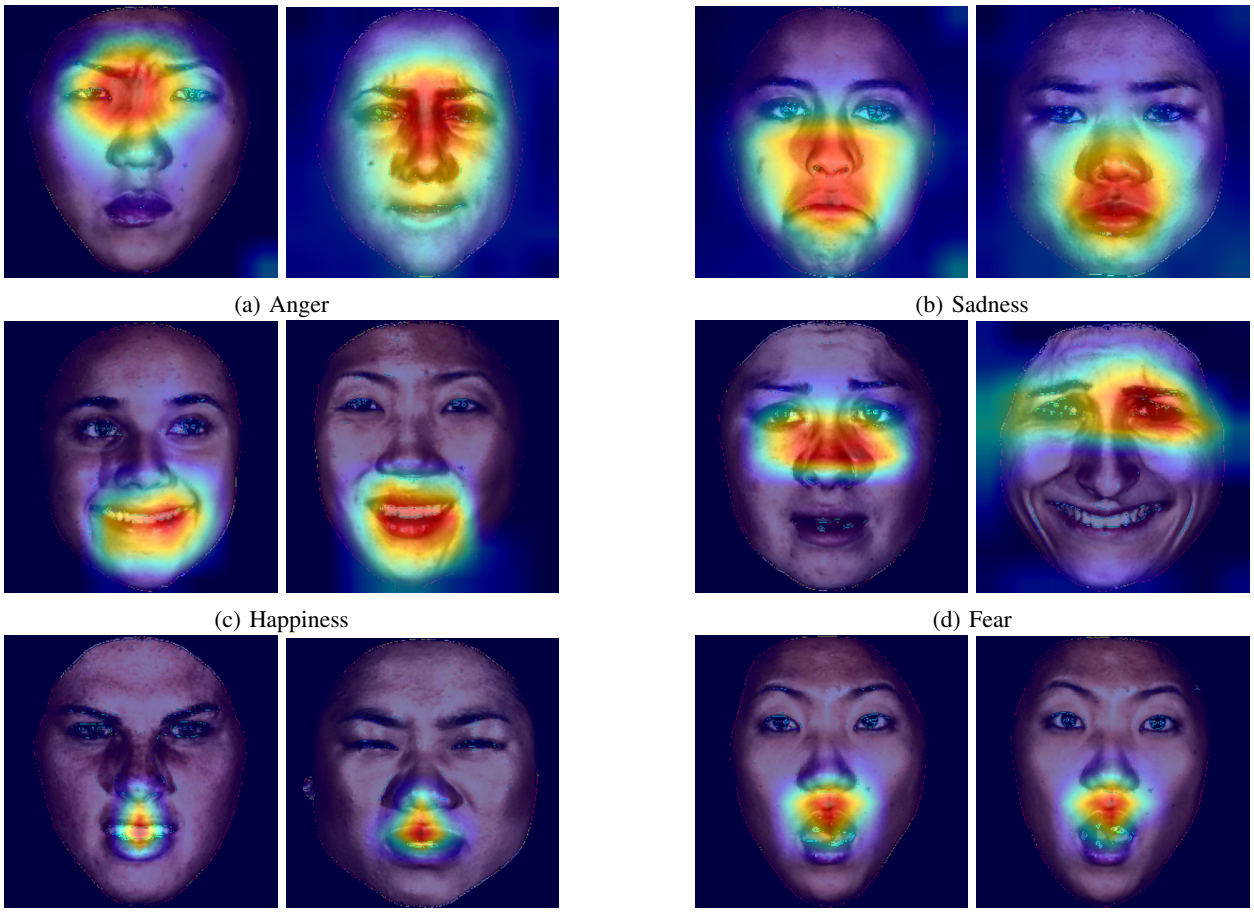
Cross Mod. Spatial Att.	ViT encoder	Accuracy(%)
✗	✗	87,25
✗	✓	88,17
✓	✗	91,45
✓	✓	91,66

Note the consistent drop in performance when not using the Cross Modality Spatial Attention, which is a good sign that the model is learning to combine the information from the two modalities in a more effective way than a sum. However, the improvement using the ViT encoder is not so relevant, which is a sign that the ViT encoder probably needs a larger dataset to be fully exploited.

D. Loss function evaluation

This section shows the discriminative power of the features at the exit of the fusion network, when trained using different loss functions. The network is trained for 10 epochs using CE Loss, Center Loss and Island Loss over CalD3rMenD3s. CalD3rMenD3s is chosen because it is a spontaneous dataset and is suitable to show the potential of the Island Loss in a real-world scenario where expressions may be more subtle and difficult to distinguish.

Figure 14 shows the features of the validation set, plot in 2D space using the t-SNE dimensionality reduction. Note that the features produced by the network trained with Center Loss are more clustered around each class center compared to the CE Loss, especially for the Neutral class which is very spread and more difficult to distinguish. Note that the results from confusion matrices are confirmed by the overlapping between classes Surprise-Fear. The Island Loss seems to provide a good clustering of the classes with better separation, even if it still struggles in the overlapping



(a) Anger (b) Sadness

(c) Happiness (d) Fear

(e) Disgust (f) Surprise

Fig. 13: Heatmaps for each emotion class

couples. The final accuracy of the model reported in Table VIII for the three cases, confirms this observation, with the Island Loss providing the best performance, followed by the Center Loss and the CE Loss.

TABLE VIII: Loss experiments

Loss	Accuracy (%)
CE	59,81
Center	61,73
Island	62,50

E. Dataset Merge

As an additional experiment, CalD3rMenD3s, BU3DFE and Bosphorus datasets are merged into a single dataset for a final testing of the model. Whether it is true that CalD3rMenD3s is spontaneous while BU3DFE and Bosphorus are posed, the low intensity expressions from BU3DFE can be assimilated as spontaneous expression, while the overly exaggerated expressions both from BU3DFE and Bosphorus can be useful for the model to capture the most important features of the expressions and transfer that knowledge over less intense expressions.

The final dataset is composed of 12098 images which distribution is shown in Figure 15 which is then split into a training set and a test set using 80%-20% policy. Final results are shown in Table IX with confusion matrix and training statistics in Figures 17 and 16

TABLE IX: Results on Global dataset

Model	Year	Accuracy (%)	Classes
This	2024	68,48	7

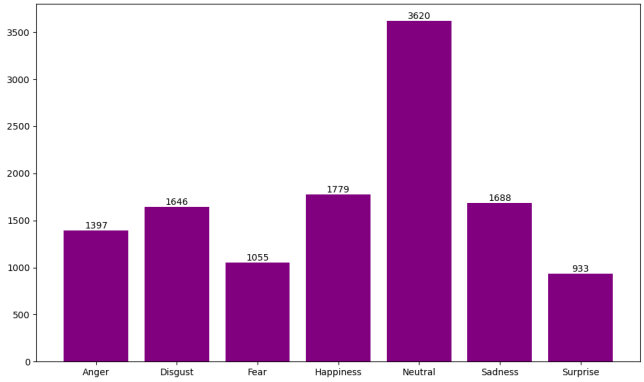
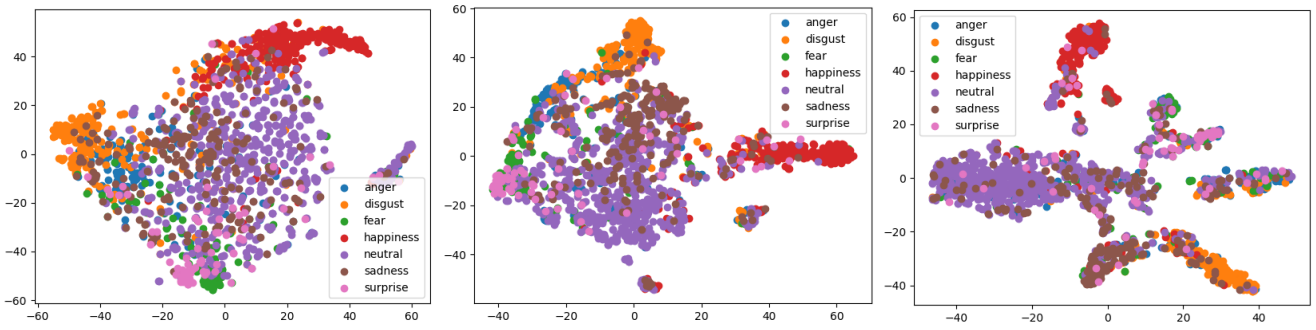
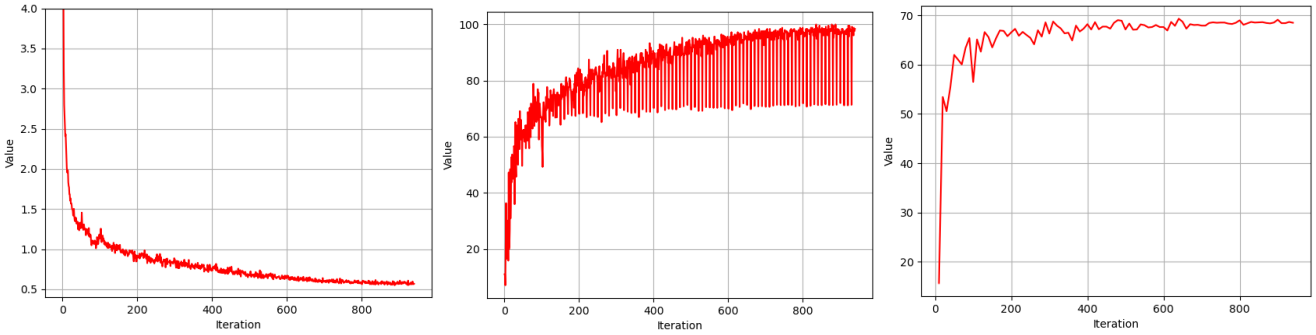


Fig. 15: Classes distribution of the Global dataset



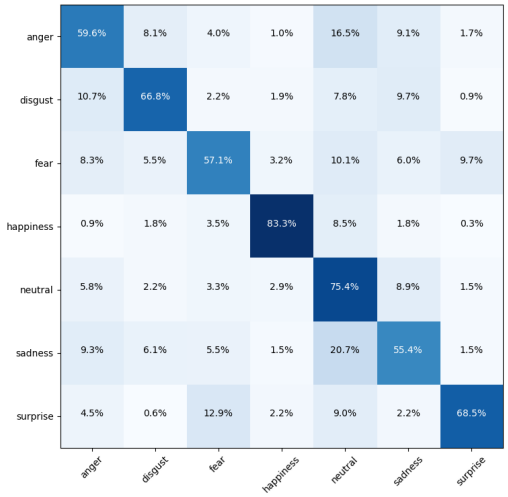
(a) Cross Entropy Loss (b) Center Loss (c) Island Loss

Fig. 14: t-SNE plot features of validation samples for different loss functions 10 epochs training.

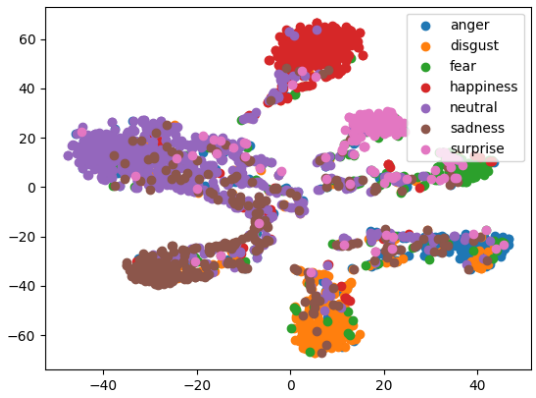


(a) Training Loss (b) Training Accuracy (c) Validation Accuracy

Fig. 16: Training and Validation Statistics for the Global dataset 7 classes



(a) Confusion Matrix. X-axis: predicted class, Y-axis: true class



(b) Features

Fig. 17: Training results on Global dataset 7 classes

V. CONCLUSIONS

This work addressed 2D+3D Facial Emotion Recognition (FER) with a hybrid deep network, which uses features extracted from a Convolutional Neural Network (CNN) to train a Vision Transformer Encoder. The proposed network was evaluated in various experimental setups over the CalD3r, MenD3s, BU3DFE, and Bosphorus datasets, along with a new multimodal dataset formed by merging these three datasets. It demonstrated superior performance across the evaluated benchmarks, achieving an accuracy of 91,66% on BU3DFE, 62,50% on CalD3r and MenD3s and 68,48% on the merged dataset. Notably, this work introduces a new benchmark on the BU3DFE dataset by using the entire dataset, providing a more comprehensive evaluation for future research. Additionally, a new multimodal dataset—created by merging the CalD3r, MenD3s, BU3DFE, and Bosphorus datasets—is proposed. This new dataset represents the largest static multimodal dataset for FER to date, combining RGB and depth modalities to better capture geometric information and improve robustness to challenging conditions such as occlusions, lighting variations, and identity-related factors.

The strength of the model lies in actively selecting the most meaningful spatial regions from the features extracted by the CNN, followed by the transformer, which learns a more global representation of the image. Future work could explore introducing spatial attention directly within the backbone networks, rather than only at the final stage, creating

an "overloaded" Squeeze-Excite module. This module could apply attention not only over the channel dimension but also across spatial dimensions. Such attention mechanisms could be applied independently in each modality or in a cross-modality setup. This concept could also be implemented in various ways, such as by rotating input features as seen in CBAM.

To further improve model generalization, expanding the dataset with additional resources, such as the 4DFAB dataset, could be beneficial. Larger datasets could also support the use of deeper and larger backbones, potentially improving model performance. Additionally, the model could be adapted to address annotation ambiguity by incorporating a text encoder, as seen in OpenAI's CLIP model, or by using a variant of batch normalization that allows the network to learn to ignore label noise.

A more thorough cross-dataset analysis could be valuable, such as training the model on the global dataset and evaluating its generalization to datasets like 4DFAB.

Finally, given the model's relatively straightforward architecture, it could be integrated into real-time applications for use in real-world scenarios. A device capable of recording RGBD images in real-time could feed these inputs into the model, providing robust predictions even in cases of occlusion or poor lighting conditions

REFERENCES

[1] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *CoRR* abs/1907.10902 (2019). arXiv: 1907.10902. URL: http://arxiv.org/abs/1907.10902.

[2] Haoran Bai et al. "FFHQ-UV: Normalized Facial UV-Texture Dataset for 3D Face Reconstruction". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 362–371. URL: https://api.semanticscholar.org/CorpusID:254018271.

[3] Ankan Bansal et al. "UMDFaces: An annotated face dataset for training deep networks". In: *2017 IEEE International Joint Conference on Biometrics (IJCB)* (2016), pp. 464–473. URL: https://api.semanticscholar.org/CorpusID:66176.

[4] Emad Barsoum et al. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution". In: *CoRR* abs/1608.01041 (2016). arXiv: 1608.01041. URL: http://arxiv.org/abs/1608.01041.

[5] Margaret M. Bradley and Peter J. Lang. "International Affective Picture System". In: *Encyclopedia of Personality and Individual Differences*. Ed. by Virgil Zeigler-Hill and Todd K. Shackelford. Cham: Springer International Publishing, 2017, pp. 1–4. ISBN: 978-3-319-28099-8. DOI: 10.1007/978-3-319-28099-8_42-1. URL: https://doi.org/10.1007/978-3-319-28099-8_42-1.

[6] Jie Cai et al. "Island Loss for Learning Discriminative Features in Facial Expression Recognition". In: *CoRR* abs/1710.03144 (2017). arXiv: 1710.03144. URL: http://arxiv.org/abs/1710.03144.

[7] Qiong Cao et al. "VGGFace2: A dataset for recognising faces across pose and age". In: *CoRR* abs/1710.08092 (2017). arXiv: 1710.08092. URL: http://arxiv.org/abs/1710.08092.

[8] Sheng Chen et al. "MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices". In: *CoRR* abs/1804.07573 (2018). arXiv: 1804.07573. URL: http://arxiv.org/abs/1804.07573.

[9] Shiyang Cheng et al. "4DFAB: A Large Scale 4D Facial Expression Database for Biometric Applications". In: *CoRR* abs/1712.01443 (2017). arXiv: 1712.01443. URL: http://arxiv.org/abs/1712.01443.

[10] Elise S. Dan-Glauser and Klaus R. Scherer. "The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance". In: *Behavior Research Methods* 43 (2011), pp. 468–477. URL: https://api.semanticscholar.org/CorpusID:207655542.

[11] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *CoRR* abs/1801.07698 (2018). arXiv: 1801.07698. URL: http://arxiv.org/abs/1801.07698.

[12] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929. URL: https://arxiv.org/abs/2010.11929.

[13] Paul Ekman and W V Friesen. "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17 2 (1971), pp. 124–9. URL: https://api.semanticscholar.org/CorpusID:14013552.

[14] Paul Ekman and Wallace V. Friesen. "Facial Action Coding System: Manual". In: 1978. URL: https://api.semanticscholar.org/CorpusID:140895661.

[15] Dongyoon Han et al. "ReXNet: Diminishing Representational Bottleneck on Convolutional Neural Network". In: *CoRR* abs/2007.00992 (2020). arXiv: 2007.00992. URL: https://arxiv.org/abs/2007.00992.

[16] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[17] Kaiming He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *CoRR* abs/2111.06377 (2021). arXiv: 2111.06377. URL: https://arxiv.org/abs/2111.06377.

[18] Myung Beom Her et al. *Batch Transformer: Look for Attention in Batch*. 2024. arXiv: 2407.04218 [cs.CV]. URL: https://arxiv.org/abs/2407.04218.

[19] Andrew G. Howard et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *ArXiv* abs/1704.04861 (2017). URL: https://api.semanticscholar.org/CorpusID:12670695.

[20] Gary B. Huang et al. "Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments". In: 2008. URL: https://api.semanticscholar.org/CorpusID:88166.

[21] Qionghao Huang et al. "Facial expression recognition with grid-wise attention and visual transformer". In: *Information Sciences* 580 (2021), pp. 35–54. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2021.08.043. URL: https://www.sciencedirect.com/science/article/pii/S0020025521008495.

[22] Rachael E. Jack et al. "Facial expressions of emotion are not culturally universal". In: *Proceedings of the National Academy of Sciences* 109 (2012), pp. 7241–7244. URL: https://api.semanticscholar.org/CorpusID:2661203.

[23] Prannay Khosla et al. "Supervised Contrastive Learning". In: *CoRR* abs/2004.11362 (2020). arXiv: 2004.11362. URL: https://arxiv.org/abs/2004.11362.

[24] Hanting Li et al. "MFEViT: A Robust Lightweight Transformer-based Network for Multimodal 2D+3D Facial Expression Recognition". In: *CoRR* abs/2109.13086 (2021). arXiv: 2109.13086. URL: https://arxiv.org/abs/2109.13086.

[25] Shan Li, Weihong Deng, and JunPing Du. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2584–2593. DOI: 10.1109/CVPR.2017.277.

[26] Yande Li et al. *FER-former: Multi-modal Transformer for Facial Expression Recognition*. 2023. arXiv: 2303.12997 [cs.CV]. URL: https://arxiv.org/abs/2303.12997.

[27] Shisong Lin et al. "Orthogonalization-Guided Feature Fusion Network for Multimodal 2D+3D Facial Expression Recognition". In: *IEEE Transactions on Multimedia* 23 (2020), pp. 1581–1591. URL: https://api.semanticscholar.org/CorpusID:225718070.

[28] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.

[29] Patrick Lucey et al. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (2010), pp. 94–101. URL: https://api.semanticscholar.org/CorpusID:3329621.

[30] Fuyan Ma, Bin Sun, and Shutao Li. "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion". In: *IEEE Transactions on Affective Computing* 14.2 (2023), pp. 1236–1248. DOI: 10.1109/TAFFC.2021.3122146.

[31] Jiawei Mao et al. *POSTER++: A simpler and stronger facial expression recognition network*. 2023. arXiv: 2301.12149 [cs.CV]. URL: https://arxiv.org/abs/2301.12149.

[32] Daniel Miller, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. "MegaFace: A Million Faces for Recognition at Scale". In: *ArXiv* abs/1505.02108 (2015). URL: https://api.semanticscholar.org/CorpusID:15074951.

[33] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *CoRR* abs/1708.03985 (2017). arXiv: 1708.03985. URL: http://arxiv.org/abs/1708.03985.

[34] Rongrong Ni et al. "Facial Expression Recognition Through Cross-Modality Attention Fusion". In: *IEEE Transactions on Cognitive and Developmental Systems* 15.1 (2023), pp. 175–185. DOI: 10.1109/TCDS.2022.3150019.

[35] Mang Ning, Albert Ali Salah, and Itir Onal Ertugrul. *Representation Learning and Identity Adversarial Training for Facial Behavior Understanding*. 2024. arXiv: 2407.11243 [cs.CV].

[36] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[37] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *CoRR* abs/1409.0575 (2014). arXiv: 1409.0575. URL: http://arxiv.org/abs/1409.0575.

[38] A. Savchenko, Lyudmila V. Savchenko, and Ilya Makarov. "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network". In: *IEEE Transactions on Affective Computing* 13 (2022), pp. 2132–2143. URL: https://api.semanticscholar.org/CorpusID:250298227.

[39] Andrey V. Savchenko. "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks". In: *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. 2021, pp. 119–124. DOI: 10.1109/SISY52375.2021.9582508.

[40] Arman Savran et al. "Bosphorus Database for 3D Face Analysis". In: Jan. 2008, pp. 47–56. ISBN: 978-3-540-

89990-7. DOI: 10.1007/978-3-540-89991-4_6.

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *CoRR* abs/1503.03832 (2015). arXiv: 1503.03832. URL: http://arxiv.org/abs/1503.03832.

[42] Christoph Schuhmann et al. "LAION-5B: An open large-scale dataset for training next generation image-text models". In: *ArXiv* abs/2210.08402 (2022). URL: https://api.semanticscholar.org/CorpusID:252917726.

[43] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[44] Leslie N. Smith and Nicholay Topin. "Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates". In: *CoRR* abs/1708.07120 (2017). arXiv: 1708.07120. URL: http://arxiv.org/abs/1708.07120.

[45] Ming-Fa Sui et al. "FFNet-M: Feature Fusion Network with Masks for Multimodal Facial Expression Recognition". In: *2021 IEEE International Conference on Multimedia and Expo (ICME)* (2021), pp. 1–6. URL: https://api.semanticscholar.org/CorpusID:236227917.

[46] Mingzhe Sui et al. "AFNet-M: Adaptive Fusion Network with Masks for 2D+3D Facial Expression Recognition". In: *2023 IEEE International Conference on Image Processing (ICIP)*. 2023, pp. 116–120. DOI: 10.1109/ICIP49359.2023.10222441.

[47] Christian Szegedy et al. "Going Deeper with Convolutions". In: *CoRR* abs/1409.4842 (2014). arXiv: 1409.4842. URL: http://arxiv.org/abs/1409.4842.

[48] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *CoRR* abs/1905.11946 (2019). arXiv: 1905.11946. URL: http://arxiv.org/abs/1905.11946.

[49] Onkar Sumant Tanmay Sharma. In: *Emotion Detection and Recognition Market Size, Share, Competitive Landscape and Trend Analysis Report by Software Tool, by Application, by Technology, by End User : Global Opportunity Analysis and Industry Forecast, 2021-2031*. 2023, p. 232. URL: https://www.alliedmarketresearch.com/emotion-detection-and-recognition-market.

[50] Luca Ulrich et al. "CalD3r and MenD3s: Spontaneous 3D facial expression databases". In: *Journal of Visual Communication and Image Representation* 98 (2024), p. 104033. ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2023.104033. URL: https://www.sciencedirect.com/science/article/pii/S1047320323002833.

[51] Qiangchang Wang and Guodong Guo. "LS-CNN: Characterizing Local Patches at Multiple Scales for Face Recognition". In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 1640–1653. DOI: 10.1109/TIFS.2019.2946938.

[52] Azmine Toushik Wasi et al. *ARBEx: Attentive Feature Extraction with Reliability Balancing for Robust Facial Expression Learning*. 2023. arXiv: 2305.01486 [cs.CV]. URL: https://arxiv.org/abs/2305.01486.

[53] Yandong Wen et al. "A Discriminative Feature Learning Approach for Deep Face Recognition". In: *European Conference on Computer Vision*. 2016. URL: https://api.semanticscholar.org/CorpusID:4711865.

[54] Lior Wolf, Tal Hassner, and Itay Maoz. "Face recognition in unconstrained videos with matched background similarity". In: *CVPR 2011*. 2011, pp. 529–534. DOI: 10.1109/CVPR.2011.5995566.

[55] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module". In: *CoRR* abs/1807.06521 (2018). arXiv: 1807.06521. URL: http://arxiv.org/abs/1807.06521.

[56] Fanglei Xue, Qiangchang Wang, and Guodong Guo. "TransFER: Learning Relation-aware Facial Expression Representations with Transformers". In: *CoRR* abs/2108.11116 (2021). arXiv: 2108.11116. URL: https://arxiv.org/abs/2108.11116.

[57] Dong Yi et al. "Learning Face Representation from Scratch". In: *ArXiv* abs/1411.7923 (2014). URL: https://api.semanticscholar.org/CorpusID:17188384.

[58] Lijun Yin et al. "A 3D facial expression database for facial behavior research". In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 2006, pp. 211–216. DOI: 10.1109/FGR.2006.6.

[59] Kaihao Zhang et al. "EDFace-Celeb-1M: Benchmarking Face Hallucination With a Million-Scale Dataset". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2021), pp. 3968–3978. URL: https://api.semanticscholar.org/CorpusID:238583342.

[60] Saining Zhang et al. "A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition". In: *Electronics* 12.17 (2023). ISSN: 2079-9292. DOI: 10.3390/electronics12173595. URL: https://www.mdpi.com/2079-9292/12/17/3595.

[61] Zhaoqing Zhu et al. "CMANET: Curvature-Aware Soft Mask Guided Attention Fusion Network for 2D+3D Facial Expression Recognition". In: *2022 IEEE International Conference on Multimedia and Expo (ICME)* (2022), pp. 1–6. URL: https://

api . semanticscholar . org / CorpusID :
251847634.