Advanced Facial Emotion Classification with 135 Classes for Enhanced Cybersecurity Applications

1st Gregory Powers
Mathematics and Computer Science
Wabash College
Crawfordsville, U.S.
gapowers27@wabash.edu

2nd Aobo Jin
Computer Science
University of Houston-Victoria
Victoria, U.S.
JinA@uhv.edu

3rd Abdul Basit Tonmoy
Mathematics and Computer Science
Wabash College
Crawfordsville, U.S.
atonmoy27@wabash.edu

4th Haowei Cao
Electrical and Computer Engineering
Johns Hopkins University
Baltimore, U.S.
hcao28@jh.edu

5th Hardik Gohel
Computer Science
University of Houston-Victoria
Victoria, U.S.
gohelh@uhv.edu

6th Qixin Deng*
Mathematics and Computer Science
Wabash College
Crawfordsville, U.S.
dengq@wabash.edu

***Abstract*—A key component of behavior analysis and human-computer interaction (HCI) is facial expression detection, which helps systems understand and react to human emotions more effectively and nuancedly.While previous research predominantly focuses on classifying seven or eight primary emotions, there has been limited exploration into more detailed and varied emotional expressions. In this paper, we address this gap by utilizing a recently published dataset that includes 135 distinct emotional classes, significantly increasing the complexity and granularity of emotion recognition tasks. To tackle this challenge, we propose a novel framework that combines U-Net and ResNet architectures, leveraging the strengths of each to capture intricate emotional cues and spatial features. Our approach achieves higher accuracy than current state-of-the-art (SOTA) methods, demonstrating its effectiveness for high-dimensional emotion classification in complex, real-world applications.**

***Index Terms*—facial emotion recognition, deep learning, computer vision**

I. Introduction

The advancement of behavior analysis and human-computer interaction (HCI) depends on facial emotion recognition (FER), which helps systems better understand and react to human emotions [1], [2]. By allowing computational systems to understand human affective states, FER enhances applications ranging from robotics and virtual reality environments [3] to psychological assessments [4]. The significant impact of accurately recognizing human emotions underscores the necessity for robust FER systems capable of navigating the complexities of human affective expressions.

Despite substantial progress, most existing research focuses on classifying a limited set of primary emotions—typically seven or eight categories such as happiness, sadness, anger, fear, disgust, surprise, and neutrality [5]–[7]. This emphasis on basic emotions, while foundational, restricts the depth and applicability of FER systems in real-world scenarios where emotional expressions are far more nuanced [8]. Human emotions often involve subtle variations and combinations that convey a richer spectrum of affective experiences, including complex emotions like frustration, embarrassment, or skepticism. Limiting FER systems to primary emotions hinders their potential to fully understand and engage with the intricacies of human emotional expression.

To address the limitations of traditional FER methods, researchers have developed several annotated datasets that enhance the complexity and richness of emotional expressions. Some datasets combine basic emotions to form compound expressions, capturing a broader range of human affect [8], [9]. Others move away from discrete labels and adopt multi-label distributions to represent the intensity and coexistence of multiple emotions [10]. Additionally, efforts have been made to expand the set of emotion categories by including more nuanced classes [11], [12]. Some researchers use continuous dimensions, including valence (positive to negative) and arousal (calm to aroused), to express emotions rather than discrete labels. The intricacy of human emotions and subjective interpretations make it difficult to appropriately annotate continuous labels, even if this dimensional method provides a more sophisticated knowledge of emotional states [6]. Despite these advancements, the issue of semantic richness in recognizable emotion concepts remains unresolved, presenting a significant challenge to the FER community.

Recent research has focused on the granularity of emotion concepts in order to address the semantic richness problem in FER. Chen et al. [13] chose 135 emotion names among hundreds of English emotion phrases, drawing influence from psycho-linguistic studies. By gathering a sizable 135-class FER image collection, they examined the related facial expressions and put forth a corresponding framework for facial emotion identification. This dataset provides an unprecedented opportunity to explore detailed FER, but the high dimensionality introduces substantial challenges. These include increased computational demands and the need for sophisticated feature extraction and classification techniques capable of handling

979-8-3315-1888-2/25/$31.00

intricate emotional cues.

Thanks to recent advancements in deep learning, facial emotion recognition (FER) has been significantly enhanced by enabling automatic feature extraction and robust classification. Convolutional Neural Networks (CNNs) have become fundamental in FER due to their ability to learn hierarchical facial features directly from raw images [2]. To address challenges such as occlusions and variations in pose and lighting, researchers have incorporated attention mechanisms and residual connections into deep learning models. Wang et al. [14] proposed Region Attention Networks that focus on crucial facial regions, improving recognition accuracy under occluded conditions. Transformer-based architectures have also been introduced for FER; Huang et al. [15] leveraged self-attention mechanisms to effectively model relationships between different facial regions.

We propose a new framework that combines the advantages of the ResNet [16] and U-Net [17], [18] architectures. The U-Net architecture is well known for its efficacy in picture segmentation tasks because it uses a symmetric expanding path and a contracting path to capture spatial hierarchy and context. By allowing the creation of extremely deep networks through residual learning, which alleviates problems with vanishing gradients, ResNet, on the other hand, excels in picture categorization. By integrating these two architectures, our framework leverages U-Net's capability to preserve spatial information and ResNet's proficiency in deep feature representation. This feature allows for the capture of intricate emotional cues and spatial features essential for distinguishing among a large number of emotional classes.

Our approach demonstrates superior performance compared to current state-of-the-art (SOTA) methods [13], [19], [20], achieving higher accuracy in classifying the 135 emotional classes. This advancement not only validates the effectiveness of our framework but also signifies a meaningful step toward more nuanced and comprehensive FER systems. The improved accuracy in high-dimensional emotion classification has substantial implications for complex, real-world applications where understanding subtle emotional variations is crucial.

The contributions of this paper are listed below:

- **Effective Framework**: We propose a novel framework that combines U-Net and ResNet architectures to effectively capture intricate emotional cues and spatial features necessary for high-dimensional emotion classification.
- **Performance Enhancement**: We demonstrate that our approach achieves higher accuracy than existing SOTA methods, validating its effectiveness for complex, real-world applications and advancing the field of FER.

The remainder of this paper is organized as follows: In Section II, we review related works in facial emotion recognition and deep learning architectures close to our study. Section III introduces the dataset, including dataset info, preprocess, etc. Section **??** details the proposed framework, including the architectural design and training methodologies. Experimental results and performance evaluations are presented in Section V. Finally, Section V concludes the paper, summarizing our contributions and highlighting the significance of our work.

II. Related Works

This section reviews the progression of FER, including datasets, traditional methods, and recent advancements in deep learning, highlighting the strengths and limitations of each approach.

A. FER Dataset

a) Discrete Emotion Datasets: Early FER research relied on datasets focusing on basic emotions. A thorough method for categorizing facial movements was offered by Ekman and Friesen's Facial Action categorizing System (FACS) [21]. FER-Wild [22], the Cohn-Kanade AU-Coded Expression Database [23], and the Japanese Female Facial Expression (JAFFE) database [24] are a few of the popular datasets that include posed expressions of fundamental emotions. These datasets are well-annotated and serve as benchmarks for FER algorithms for a relatively long time, they often consist of posed expressions in controlled environments and its incapability of modeling fine-grained emotion variances.

b) Compound Emotion Datasets: To capture the complexity of human emotions, researchers developed datasets with compound emotions. Du et al. [8] created a dataset of compound facial expressions by combining basic emotions, highlighting the need for recognizing blended tags. In order to convert the discrete emotion labels into continuous distributions, Barsoum et al. [25] ask ten crowd taggers to relabel every image in the FER data set. The first database with compound expressions found in the wild is RAF-DB [26], which has 12 class compound emotions and 6 class basic emotions. The first database in the wild to offer crowdsourced cognition for multi-label expressions is RAF-ML [27].

c) Semantic-rich Emotion Datasets: More than 200k samples and 54 distinct fine-grained expression types are included in F^2ED [12], which significantly addresses the dearth of diversity in the current expression datasets. Chen et al. [13] drew inspiration from psycho-linguistic research and investigated the corresponding facial expressions by collecting a large-scale 135-class FER image dataset.

B. Facial Emotion Classification and Recognition

a) Traditional Method: Early FER systems relied on handcrafted features and traditional machine learning algorithms. Lyons et al. [24] used Gabor wavelets to extract facial features due to their ability to capture spatial frequency and orientation information. The method is kind of effective in capturing local facial features, but Computation load is intensive and sensitive to variations in scale and rotation. [28] employed LBP for texture description in facial images. It is more computationally efficient and robust to illumination changes, but limited in capturing global facial structure and high-level features. Carcagnì et al. [29] utilized Histograms of Oriented Gradients (HOG) features for FER. The proposed method is effective for capturing edge information and facial

contours but may not capture fine-grained facial details necessary for nuanced emotion recognition.

b) Deep Learning in FER: The advent of deep learning has revolutionized FER by enabling automatic feature extraction from raw images. Convolutional Neural Networks (CNNs): Deep CNNs like ResNet [16] have achieved promising performance in FER. [30] proposed learning facial expression embeddings disentangled from identity, addressing the issue of identity bias. In the presence of occlusions, [31] suggested an effective gACNN to augment global face information for FER. In order to concentrate on the areas of the face where muscles move, [32] presented a Facial Motion Prior Networks (FMPN) architecture that added a branch to create a facial mask. [33] proposed an end-to-end trainable framework, GeoCNN, which improves AU detection, and presented a unique GeoConv operation that integrates 3D information into 2D convolution without introducing extra trainable parameters. [34] presented an end-to-end learning model that carries out pose-invariant facial expression recognition and face image synthesis at the same time. In order to learn robust features for FER, [35] presented a self-cure network (SCN) that is intended to lower uncertainty in facial expression data.

III. Dataset (Emo135)

The data set used, Emo135 has 135 classes, and 697,354 images, split into 557,179 training images, 70,209 images for testing, and 69,966 for validation. It includes a wide variety of images of faces with expressions correlated with specific emotions. It includes no only basic emotions such as fear, anger, or happiness, but also much more nuanced and complex emotions such as cheerfulness, guilt, or gloom. Due to the difficulty of gathering organic data, data was originally obtained via utilizing search engines to automatically find images which match certain keywords, both the name of an emotion as well as other related keywords such as ”feeling” or ”expression.” This data set was then filtered using a KNN (K Nearest Neighbors) algorithm, which removed any data which was too far of an outlier from the rest of the dataset.

It should be noted that the dataset does not contain roughly equal numbers of images for each emotion, with the numbers varying wildly. For example, ”excitement” is the largest category in the training data with 27,212 images, while ”ferocity” has the least with only 28 images. this makes high accuracy across the dataset a very difficult task to achieve. it should therefore be expected that some categories have much higher accuracy percentages than others.

A. Preprocessing

Each image is preprocessed in this manner: first calculate the bounds of a face in the image using an MTCNN model. If no face is found, discard the image. Otherwise, scale the image up until the area of the image which contains the face is smaller than the target resolution of 224x224. Alternately, if the image size is smaller than our target size, scale it up so its smallest dimension equals the target dimension. Crop the image so that the entire box outlining the face is contained in it along with some margin, and so that it becomes our target resolution.

Fig. 1. This figure shows the steps an image goes through while being preprocessed for use in training.

During preprocessing we had to discard some amount of images, leaving us with a total of 668,974 images. Of these, 534,567 are for training, 67,263 for testing, and 67,134 for validation.

IV. Model

For this project we use a combination U-Net/ResNet architecture. The U-Net is first applied to the input image in order to generate a segmentation map which in this case acts as a mask. This mask is then combined with the original image and some randomly generated noise via the formula $image = image * mask + (1 - mask) * noise$. This noise ensures that not all detail is lost from the masked areas. This image is then fed into the ResNet model, which performs the actual feature extraction and classification tasks. For our purposes, we use a 34 layer ResNet, since it seems to give the best results for the amount of data in our dataset. The intent of including a U-Net in our model architecture is that this part of the model can learn which features or areas in an image are less important to the actual classification task, e.g. the neck or ears, in favor of more important ones, e.g., the eyes or mouth. In general, we would like it to generate a segmentation map that prioritizes the features we care about, which can then be turned into a mask. The ResNet is used as the backbone of our model because of its exceptional past performance in image classification tasks.

A. ResNet

Convolutional neural networks (CNNs) have demonstrated substantial efficacy across diverse image classification tasks due to their hierarchical feature extraction capabilities. In particular, deep convolutional networks—those that stack a significant number of layers—have shown remarkable performance improvements. The underlying rationale is that as layers increase in depth, the network’s ability to capture complex, high-level abstractions also increases. Lower layers may capture basic features such as edges and textures, while deeper layers represent more abstract constructs, such as shapes, objects, and even parts of scenes. However, increasing network depth has practical limits due to optimization challenges, specifically the vanishing gradient problem, which frequently emerges as a significant impediment to training very deep architectures. The vanishing gradient problem occurs during back-propagation, where the gradients—used to adjust weights

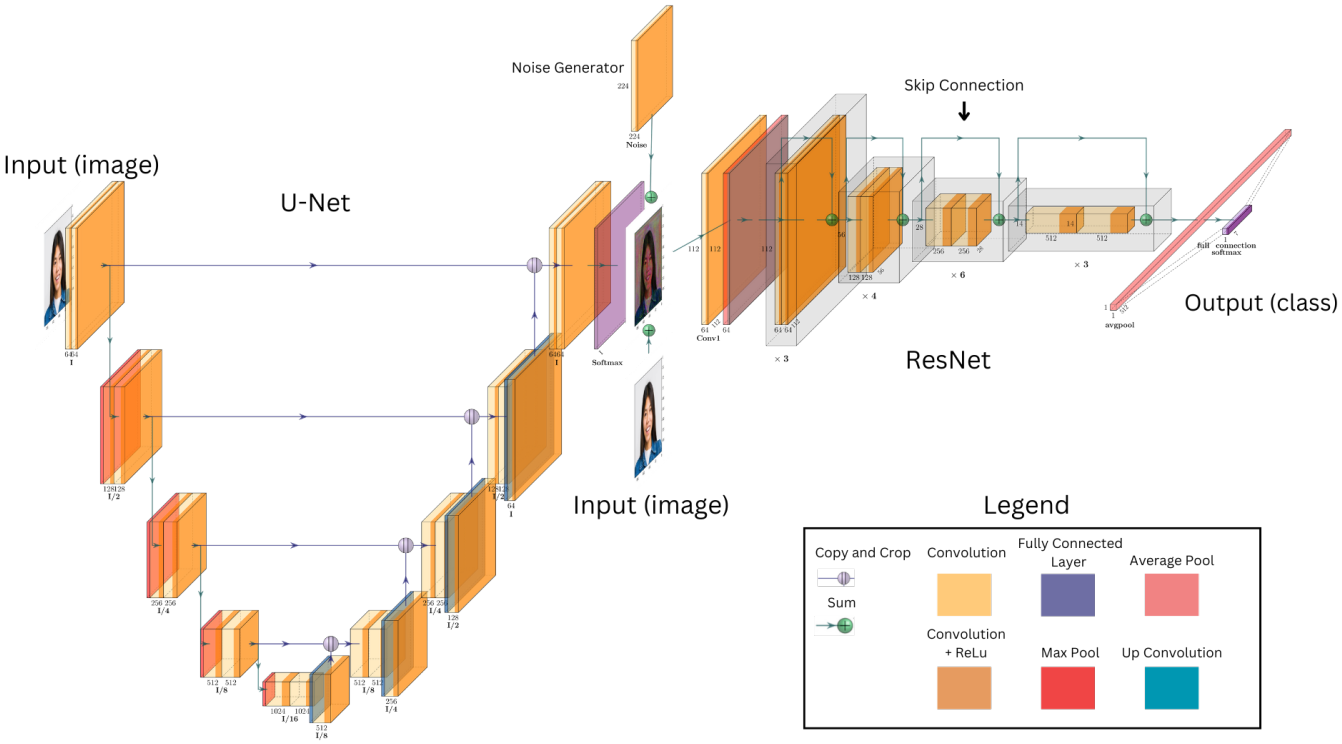

Fig. 2. An outline of the model and data flow

in order to minimize loss—become progressively smaller as they are propagated back through multiple layers. In very deep networks, this repeated multiplication of small gradient values results in near-zero gradients for early layers, which effectively halts learning in these layers. Consequently, this limits the depth of feasible CNN architectures, as deeper networks tend to suffer from degraded performance due to an inability to update weights effectively. This problem led to a critical question: how can one construct deeper networks without incurring the drawbacks associated with vanishing gradients?

Residual Networks (ResNet) represent a seminal advancement in addressing this challenge. Introduced by He et al., ResNet's core innovation is the concept of residual learning, which fundamentally redefines how each layer in a deep network learns. Instead of each layer learning a direct transformation $F(x)$ of its input x, ResNet layers are designed to learn a residual function $F(x) - x$, effectively reformulating the problem as learning the difference between the input and the target output, denoted as $F(x) = H(x) + x$, where $H(x)$ is the residual function. This adjustment is implemented through skip, or shortcut, connections, which bypass one or more layers by directly adding the input to the output of these layers. These residual connections allow the input to skip transformation layers, creating a direct pathway for the gradient during back-propagation.

This residual framework confers two major advantages. First, the skip connections allow the gradient to flow unimpeded through the network, significantly mitigating the vanishing gradient problem. With these connections, ResNet enables the training of networks that are substantially deeper (e.g., ResNet-50) without suffering from performance degradation, as gradients can bypass intermediary layers when necessary. Second, the skip connections enable the network to more easily optimize identity mappings when deeper transformations do not improve the learning objective. Consequently, ResNet layers can be seen as dynamic units that only modify the input when it contributes to a meaningful transformation, reducing unnecessary complexity and enabling efficient learning.

When further changes are neither required or beneficial, residual networks' ability to optimize identity mappings enables layers to stay closer to their original representations. This is their second key advantage. This feature allows the network to avoid unnecessary changes, protecting data while concentrating processing power solely on layers that significantly advance the learning goal.

Bypassing intermediate transformations and transferring the input of one layer straight to a subsequent layer, skip (or shortcut) connections allow for the optimization of identity mappings. The output of the layer defaults to the input x when the network determines that a transformation adds no value, and the residual function $H(x)$ effectively becomes zero. The network can keep information without changing it because to the skip connection, which makes learning easier.

This is important because transformations in very deep networks can occasionally introduce noise or redundancy into

the feature representations, which reduces rather than improves efficiency. ResNets lower the danger of superfluous complexity by allowing layers to maintain identity mappings. This helps avoid overfitting and guarantees that only advantageous transformations are used. By avoiding overcomplicating the learnt representations, this selective transformation not only increases training efficiency but also improves the model's generalizability to fresh data.

Additionally, ResNets can handle vanishing gradients even better thanks to identity mapping since the gradients can propagate back without being hampered by unnecessary transformations. One of the reasons ResNets perform admirably on a variety of tasks and maintain their stability as they get deeper is because of this blend of control and flexibility. ResNets' ability to map identities essentially allows them to dynamically adjust their complexity, learning only the information required to improve the learning goal without having to deal with unnecessary operations.

B. U-Net

U-Net is a well-known and incredibly successful model architecture for biomedical image segmentation that was created especially to handle the difficulties associated with sophisticated biomedical image segmentation. The architecture of U-Net uses a special, symmetrical network topology with two main paths—a contracting path and an expansive one—to learn how to recreate segmented images from raw, unsegmented data. Because of its design, U-Net can maintain contextual global features and fine-grained local detail throughout the segmentation process.

The contracting route of U-Net functions as a feature encoder and shares structural similarities with a traditional CNN. It is made up of a sequence of recurrent 3x3 convolutions, each of which is followed by rectified linear unit (ReLU) activations to introduce non-linearity and 2x2 max-pooling operations to reduce the spatial dimensions. In this method, each convolutional block captures increasingly abstract representations by using activation functions and convolving feature maps. The max-pooling layers reduce the spatial resolution of the image by progressively downsampling it and doubling the number of feature channels. This technique allows the model to depict a hierarchy of elements at various sizes by beginning with fine-grained details and progressing to deeper layers with more intricate, wide-ranging patterns. This slow reduction in spatial resolution allows the contracting route to acquire rich, multi-level feature representations, which are essential for distinguishing fine-grained tissue borders or small tumors, among other delicate and complex structures in medical imaging.

The bottleneck, or lowest resolution, is where U-Net's design mixes the most abstract feature representations. This core bottleneck layer acts as a bridge between the contracting and expansive routes by synthesizing contextual information from the full image prior to initiating the upsampling process. The bottleneck layer allows for the incorporation of global spatial information, which is necessary for accurately reconstructing complex segment boundaries during the upsampling stage. It also reduces the information loss associated with downsampling, ensuring that the network preserves the essential context required for precise segmentation.

In order to reconstitute the segmentation map, the image representation passes through the decoder, also referred to as the expanding path, after the bottleneck. The decoder systematically upsamples the feature maps to the original input dimensions. This extensive path consists of 2x2 up-convolutions, also known as transposed or deconvolutions, which progressively enhance the spatial resolution of the image. A balanced reconstruction of fine and coarse features is promoted by reducing the number of feature channels with each upsampling step. The transposed convolutions in this channel allow the network to identify structural boundaries even in the presence of partial blockage or blurry picture regions.

The usage of skip links between the contracting and expanded channels is one of U-Net's special and potent features. The up-sampled feature maps in the expansive path are concatenated with the equivalent high-resolution feature maps from the contracting path at each up-sampling level. High-resolution spatial information that would otherwise be lost as a result of the encoder's downsampling processes is preserved by this concatenation. The skip connections enable U-Net to efficiently integrate high-level contextual information with low-level details by incorporating these features, which is especially useful in biomedical applications where precise boundary delineation is required, such as segmenting overlapping cells, distinguishing between closely positioned anatomical structures, or identifying tumors within dense tissues

The final step in U-Net's architecture involves a 1x1 convolution layer, which reduces the number of feature channels to the desired number of segmentation classes. This output layer effectively assigns each pixel in the image to a specific class, producing a complete segmentation map that delineates each identified structure. In biomedical segmentation tasks, this output could correspond to specific regions like organs, lesions, or other anatomical structures, each represented as a separate segment with clear boundaries.

Though U-Net was initially designed for biomedical image segmentation, its architecture has demonstrated robustness and versatility across a wide range of image segmentation tasks beyond the medical domain. U-Net's success is attributed to its ability to capture multi-scale features, its effective use of spatial context preservation through skip connections, and its relatively efficient design, which requires fewer computational resources compared to many other dense architectures. U-Net has found applications in diverse fields, including remote sensing, where it segments satellite images to identify land use and cover types; autonomous driving, where it helps detect road signs, lanes, and obstacles; and natural scene understanding, where it segments elements like buildings, vegetation, and water bodies. Additionally, U-Net's architecture has inspired the development of numerous variants tailored to address specific segmentation challenges, such as attention

U-Nets, residual U-Nets, and U-Nets with 3D convolutions for volumetric data, further enhancing its applicability across different domains.

C. Training

In order to train the model, we use a stochastic gradient descent with momentum optimizer with a training rate of 0.01 and a momentum of 0.9.

The loss function chosen for this task is Cross-Entropy loss, which can be expressed by this formula:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$

Where N is the number of elements in a batch, C is the total number of classes, $y_{i,c}$ is a one hot encoded tensor of the actual correct class, and $\hat{y}_{i,c}$ is a tensor of the predicted values.

It is recommended to train the model for around 20 epochs, although it should reach peak performance well before then.

V. Experiments and Results

We test our model in comparison to other models used in image classification tasks. Each of these models is evaluated on the testing section of the previously mentioned Emo135 dataset. All training takes place on an Nvidia RTX 4090. Each model is trained until validation accuracy begins to drop off indicating overfitting. Unless otherwise mentioned, each model is trained using Stochastic Gradient descent with a learning rate of 0.01 and a momentum of 0.09, and the loss function is Cross-Entropy loss. During the training process, each epoch is saved individually, and each of these is tested on the testing dataset. The best out of these is taken to be the final accuracy.

In order to compare other methods of performing this task We swap out the backbone of our model while maintaining the U-Net component. The two alternate backbones we use are both the shallower ResNet18, as well as A base level Vision Transformer model. Vision transformers are very popular alternatives to convolutional networks for image classification tasks, and it is important to see if they give potentially better results. We also compare our model to the former state of the art model presented in *Semantic-rich facial emotional expression recognition* [13], using the result presented in that paper. In order to ascertain how our choices of optimizer and loss function are affecting the accuracy of this model, we switch out each of them and run a test to see what kind of accuracy results. We test switching out the Stochastic Gradient Descent optimizer for the Adam optimizer, and we test switching out Cross Entropy Loss for Mean Square Error.

Our model has a slight improvement of around 1% Top-! accuracy over the previous state of the art model. It clearly wins out over all the other models tested, showing that the very specific way in which it is organized is important to maintaining a high level of accuracy.

From Fig. 3 it can be seen that accuracy starts out increasing very rapidly from epoch to epoch, and then very quickly starts to drop again. This model converges very rapidly and does not benefit from large numbers of passes over the data. It is also

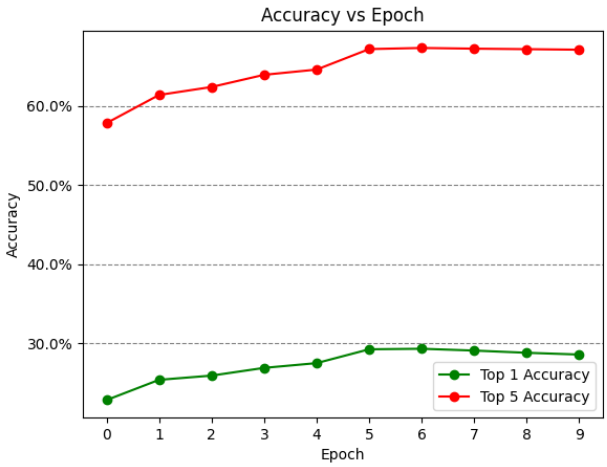

Fig. 3. This charts the accuracy of our model from epoch to epoch.

apparent that the performance of the model is also very heavily based on choice of optimizer and loss function, as switching them out leads to massively degraded performance.

It is to be expected that this model does not perform equally as well on every single different class, and therefore we have created a sample of accuracy by class based on selecting every 10th class sorted by Top-1 and Top-5 accuracy in order to emphasize the massive difference between the best and worst performance. Looking at Fig. 5 and Fig. 6 it can be seen that top 1 accuracy can vary from at best well over 70% to at worst 0%, and top 5 accuracy from well over 80% to 0%.

This could be due to one of three major causes: the difficulty of learning a specific emotion, and the number of instances of that emotion present in the training, and the quality of the images present. In order to see what affect the number of images in the training data have on the accuracy in that category of the final model, we graph the accuracy of a category in relation to the number of images of that category present in the original training dataset. For a low number of images the accuracy on a category fluctuates wildly, seemingly with no correlation to the actual number of images, but the more images are present the more accuracy stabilizes and begins to have a generally increasing trend, while still fluctuating significantly. This is quite inconclusive, but it seems that while number of images in the training dataset has a noticeable effect on accuracy, the actual difficulty of learning from images of a specific accuracy and the quality of the images have a greater effect.

VI. Discussion and Conclusion

In this study, we addressed the underexplored area of detailed facial emotion recognition (FER) by utilizing a dataset comprising 135 distinct emotional classes. Our proposed framework, which combines U-Net and ResNet architectures, demonstrates superior performance over existing state-of-the-art (SOTA) methods in high-dimensional emotion classification tasks. This advancement signifies a meaningful step

Model	Top 1 Accuracy	Top 5 Accuracy
U-Net/ResNet34[Ours]	**29.3%**	**67.3%**
U-Net/ResNet34 (Adam Optimizer)	5.00%	17.7%
U-Net/ResNet34 (MSE Loss)	0.37%	3.6%
U-Net/ResNet18	4.7%	15.1%
U-Net/Vision Transformer	4.9%	18.4%
Chen et al. [13]	28.3%	66.4%
Resnet50 + MLP	9.6%	30.6%
VGGFace2 + MLP	9.9%	32.2%
ARM [19]	20.5%	55.9%
DACL [20]	18.3%	47.7%

TABLE I
ACCURACIES OF MODELS FOR EXPRESSION CLASSIFICATION, INCLUDING OUR MODEL, MODIFIED VERSIONS OF OUR MODEL, AND MODELS FROM THE PREVIOUS SOTA PAPER.

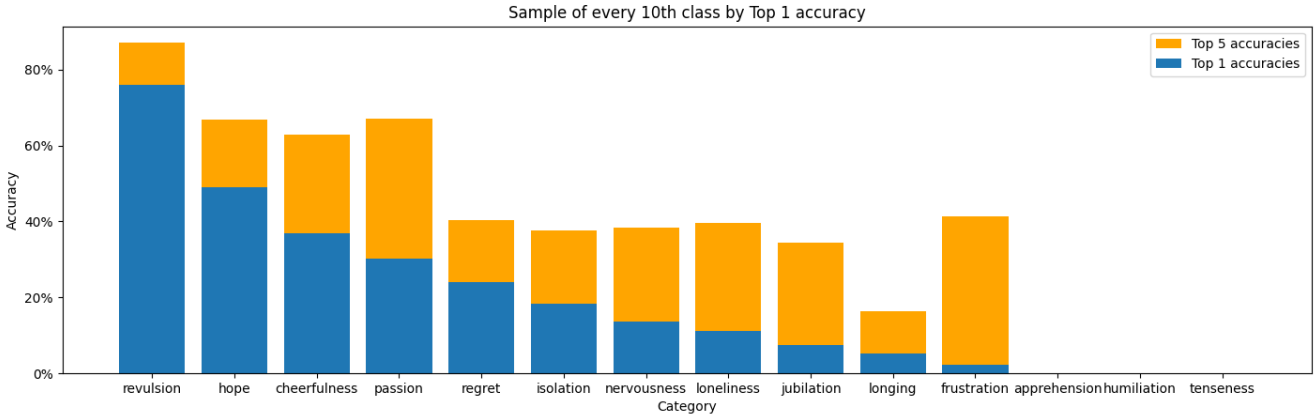

Fig. 4. A sample of the accuracy of every 10th class sorted by Top-1 accuracy, which shows the large disparity in accuracy between the best and worst performing classes. The last three classes have actually 0 accuracy in not only Top 1 but also Top 5.

toward more nuanced and comprehensive FER systems capable of handling the complexity and granularity of human emotional expressions.

Our experimental results indicate that the proposed framework outperforms current SOTA methods in both accuracy and robustness. The high accuracy achieved in classifying 135 emotional classes demonstrates the framework's capability to handle the semantic richness and granularity of emotions, which has been a significant challenge in the FER community. This performance improvement has substantial implications for real-world applications where understanding subtle emotional variations is crucial, such as in advanced human-computer interaction (HCI), social robotics, psychological assessment, and adaptive learning systems.

Despite the promising results, there is a limitation to our study that is worth discussion. Annotating 135 distinct emotional classes poses challenges due to the subjective nature of emotions and the potential for annotator inconsistency. While we have taken steps to ensure the credibility of the dataset, discrepancies in labeling could impact the model's performance. Advanced annotation techniques and consensus-

Fig. 5. Two examples of annotation inconsistency from the dataset are presented. The first row demonstrates a case where two images belong to different classes but appear very similar. In contrast, the second row showcases a case where images belong to the same class but exhibit significantly different appearances.

building methods may be necessary to improve label accuracy.

REFERENCES

[1] T. Zhang, "Facial expression recognition based on deep learning: a survey," in *Advances in Intelligent Systems and Interactive Applications: Proceedings of the 2nd International Conference on Intelligent and Interactive Systems and Applications (IISA2017)*. Springer, 2018, pp. 345–352.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[3] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 505–523, 2018.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[5] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*. Springer, 2013, pp. 117–124.

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

[7] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 279–283.

[8] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[9] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570.

[10] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 884–906, 2019.

[11] F. Zhou, S. Kong, C. C. Fowlkes, T. Chen, and B. Lei, "Fine-grained facial expression analysis using dimensional emotion model," *Neurocomputing*, vol. 392, pp. 38–49, 2020.

[12] W. Wang, Y. Fu, Q. Sun, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, Y.-G. Jiang, and X. Xue, "Learning to augment expressions for few-shot fine-grained facial expression recognition," *arXiv preprint arXiv:2001.06144*, 2020.

[13] K. Chen, X. Yang, C. Fan, W. Zhang, and Y. Ding, "Semantic-rich facial emotional expression recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1906–1916, 2022.

[14] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[15] Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Information Sciences*, vol. 580, pp. 35–54, 2021.

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

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[18] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[19] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via de-albino and affinity," *arXiv preprint arXiv:2103.10189*, 2021.

[20] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2402–2411.

[21] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[22] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 58–65.

[23] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*. IEEE, 2000, pp. 46–53.

[24] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[25] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," 2016. [Online]. Available: https://arxiv.org/abs/1608.01041

[26] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[27] ——, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vision*, vol. 127, no. 6–7, p. 884–906, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s11263-018-1131-1

[28] X. Feng, "Facial expression recognition based on local binary patterns and coarse-to-fine classification," in *The Fourth International Conference onComputer and Information Technology, 2004. CIT '04.*, 2004, pp. 178–183.

[29] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 645, 2015.

[30] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6759–6768.

[31] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[32] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," in *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2019, pp. 1–4.

[33] Y. Chen, G. Song, Z. Shao, J. Cai, T.-J. Cham, and J. Zheng, "Geoconv: Geodesic guided convolution for facial action unit recognition," *Pattern Recognition*, vol. 122, p. 108355, 2022.

[34] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3359–3368.

[35] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6897–6906.