Fine Tuning Vision Transformer Model for Facial Emotion Recognition: Performance Analysis for Human-Machine Teaming

Conference Paper · August 2023

DOI: 10.1109/IRI58017.2023.00030

2 authors:

Sanjeev Roka
Howard University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Danda B Rawat
Howard University
524 PUBLICATIONS **8,856** CITATIONS
SEE PROFILE

2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI) | 979-8-3503-3458-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IRI58017.2023.00030

Fine Tuning Vision Transformer Model for Facial Emotion Recognition: Performance Analysis for Human-Machine Teaming

Sanjeev Roka and Danda B. Rawat
Department of Electrical Engineering and Computer Science
Howard University, Washington DC 20059, USA
sanjeev.roka@bison.howard.edu, db.rawat@ieee.org

***Abstract*—Facial Emotion Recognition (FER) has become essential in various domains, including robotic systems, affective computing, emotion-triggered intelligent agents, and human-computer interaction for human-machine teaming. Although Convolutional Neural Network (CNN)-based models were popular for facial emotion classification, Transformer-based models have shown better performance in computer vision tasks such as image classification, semantic segmentation, and object detection. In this study, we explore the performance of the Vision Transformer model on a publicly available large FER dataset called AffectNet, which provides a realistic representation of emotions "in the wild." We fine-tuned the model for the emotion classification task based on facial expressions. We achieved an accuracy of 64.48% on the Affectnet validation set, outperforming many other methods that use only transformer models. Further, we explore how they can be used for Human-Machine Teaming particularly in vehicular systems to improve driver safety, comfort, and experience.**

***Index Terms*—Facial Emotion Recognition, Fine-tuning Vision Transformer, Human Machine Teaming, AffectNet**

I. Introduction

Human-Machine Teaming is a popular domain that fosters interaction between humans and machines to achieve common goals through collaborations for joint learning and reasoning [1], [2], [3]. It encompasses various technologies, including automated robots, machine learning, and artificial intelligence. A crucial aspect of successful human-machine teaming is machines' ability to comprehend human behavior and intention, and one effective method to accomplish this is through the recognition of human emotions. By analyzing the emotional state of humans, machines can make intelligent decisions. Furthermore, among different means for humans to express their emotions, facial expression is one of the most effective and powerful factors [4]. Hence, facial Emotion Recognition (FER) has become one of the important classification components in the field of computer vision. More than 20 different human emotions have been identified [5]. However, for a FER system, six different emotions are usually considered fundamental: happiness, sadness, surprise, fear, anger, and disgust [6]. A neutral emotion is often included to address neutrality alongside these six fundamental emotions. The FER system comprises three major components: face detection, facial feature extraction, and facial expression classification [7]. First, the face must be detected from the frame which can be either a picture or a video frame. The extraction of facial features is the most crucial part of the system as the final classification of the emotion completely relies on the extracted features. There are several approaches to FER systems. Most of the traditional methods relied on handcrafted features. With the introduction to larger datasets, researchers have proposed different deep-learning approaches to extract features and develop a FER classifier. The popularity of deep learning models like CNN, has been performing really well in such tasks. However, the introduction of the transformer model in computer vision tasks is gaining rapid interest as they seem to have shown superior performance as compared to CNN-based methods when trained on large-scale datasets [8].

In this paper, we use a transformer-based model, fine-tuned on a FER dataset, and analyze its performance. We take a vision transformer model that is pre-trained on the ImageNet-21k dataset and fine-tune it to build a FER classifier. For fine-tuning, we are using the AffectNet dataset which is a large database of facial expressions in the wild [9]. Since the dataset provides a large number of training samples and contains images from the wild, it is more efficient to train the transformer model so that it performs better on real-world data. We achieved a good result when fine-tuning a pre-trained Vision Transformer model using this dataset.

We organize the rest of the paper as follows. In Section II, we review different approaches that are implemented in the FER tasks. Section III explains the proposed approach that explains the overall procedure followed for the experimentation. Section IV highlights the results obtained by following our approach and discusses them. The challenges of the proposed approach, potential solutions, and possible future directions are discussed in Section V. Finally, we conclude the paper in section VI.

II. Literature Review

Several works have been done in the field of Facial Emotion Recognition over the past few decades. In this section, we discuss some of them (mainly the ones using Affectnet and

This work is funded in part by DoD Center of Excellence in AI/ML (CoE-AIML) at Howard University under Contract W911NF-20-2-0277 with the U.S. Army Research Laboratory

979-8-3503-3458-6/23/$31.00 ©2023 IEEE
DOI 10.1109/IRI58017.2023.00030

transformer). FER tasks can be widely identified as either static image-based FER or dynamic sequence FER [4]. Since we are using a static image dataset, we will be mainly focusing on the works that fall under that category. Most of the traditional FER methods, as proposed in [10], [11], [12], [13] required a separate step for feature extraction (mostly hand-crafted) and classification. But with the revolution of deep learning algorithms in the computer vision field and larger datasets made public, most of the recent FER tasks use deep learning methods for feature extraction, recognition, and classification tasks as well [14].

Work done in [15] proposes to train the FER system in two stages: first pre-training the CNN model on face recognition task and then fine-tuning it with the Affectnet dataset. They ran the experiments for both, 7 classes and 8 classes of emotion. They used the network pre-trained on VGGFace2 and replaced its last layer with a new head of a fully connected layer with outputs equal to a number of emotion classes and softmax layers. They also used cropping and rotation as the pre-processing for images and achieved a novel performance with an accuracy of 63.03% for 8 and 66.34% for 7 emotion classes. They implemented this work to develop a novel pipeline for analyzing student behavior in an e-learning environment.

The author in [16] proposes a Graph Convolutional Network (GCN) based multi-task learning (MTL) framework to recognize facial expressions. The paper combines the expression classifier and valence-arousal regressor with a GCN-based mapping whose output is then used to extract image features and perform end-to-end training. The experiment done with the proposed approach gives the current state-of-the-art performance on the Affectnet dataset with an accuracy of 66.46%. Further, the paper [17] introduced the visual cross-corpus study using eight different corpora and proposed a visual-based end-to-end emotion recognition framework implemented using two key elements: first the backbone emotion recognition model based on VGGFace2[18], ResNet50 model that is trained in a balanced way and the other is the temporal block stacked on top of the backbone model and trained with dynamic visual emotional datasets. They fine-tuned the backbone model on the AffectNet dataset and claim to achieve an accuracy of 66.4%. Apart from approaches that use only CNN-based models, several research has been carried out to combine the CNN model with an attention mechanism. One such work is proposed in [19] where the authors combine CNN with Attention mechanism [20] (ACNN) for facial expression recognition. The research primarily focuses on partial occlusions by using the attention mechanism that can put more weight/attention on the unobstructed facial regions and identify emotion more accurately. The proposed ACNN was implemented with VGG-16 as the backbone network. The model was trained on the dataset made by combining FED-RO, RAF-DB, and Affectnet which had around 400 occluded images in total. The performance of the proposed approach on the Affectnet dataset without occlusion yielded an accuracy of 58.78%. [21] proposes a transformer-based FER method(TFE) in which, the input face image encodes the convolution feature map using the ResNet18 and then encodes the facial-expression representation using a vision transformer. To address occlusion, they decoded the encoded convolution feature maps to reconstruct the occluded facial images. The method achieves an accuracy of 63.33% on the AffectNet dataset. Another approach proposed in [22] uses a plain transformer model without any CNN models and trains a FER classifier using the hybrid dataset (merged FER-2013 [23], CK+48 [24] and AffectNet[9] to form AVFER). The author uses different augmentation techniques to balance the datasets with 20000 images in each class. Finally, a vision transformer model pre-trained on the ImageNet dataset is fine-tuned to classify eight different emotions. The trained model with the ViT-B/16/S (baseline model of ViT) as the base model yielded 54.89% on 7 emotion classes. However, this approach doesn't utilize all the available image samples of the AffectNet dataset. The AffectNet dataset has more than 200K images for training only as shown in Fig 2a. But the author in [22] uses only a small portion combined with two other datasets. In our approach, we use all of the AffectNet datasets along with data augmentation to increase the number of samples of the classes that has a comparatively very less number of samples. Further, we fine-tune the Vision Transformer model with this dataset and test on the Affectnet validation set and achieve better performance.

III. Proposed Approach

In this paper, we use a transformer-based model that is gaining popularity with its state-of-the-art performances on several computer vision benchmark tasks. In particular, we use the Vision Transformer (ViT) model [25] that is replacing the convolution neural networks (CNN) [26]. We take the pre-trained vision transformer model initially trained on the Imagenet21k dataset and fine-tune the model for the facial emotion classification task using the existing FER dataset, AffectNet. The entire approach can be explained in the following steps.

A. Data Preparation

One of the major problems with most facial emotion recognition systems is that they are trained on lab-generated datasets and perform really well on data, generated in a controlled environment. However, when it comes to real-world data, the model performs poorly. Also, the transformer models need a large number of data samples to perform well.

1) Dataset Description: For this experimentation, we are using the Affectnet dataset. AffectNet is a large database of facial expressions in the wild that are collected from the internet by querying different search engines and are annotated manually. It is by far the largest database having more than one million facial images. The annotated dataset consists of 8 different facial emotions: neutral. happy, sad, surprised, fearful, disgusted, angry, and contemptuous. For our training, we are discarding the contempt emotion. Hence, we have 283901 training samples along with 3500 validation samples [9]. Some of the sample images from the Affectnet dataset are

shown in Fig.1. Also, the class distribution for the training and validation set of the AffectNet dataset is shown in Fig 2

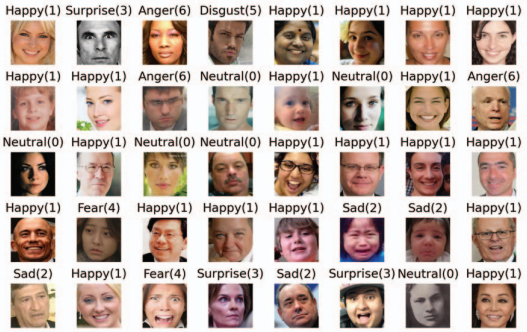


Fig. 1: Sample images of AffectNet dataset with labels

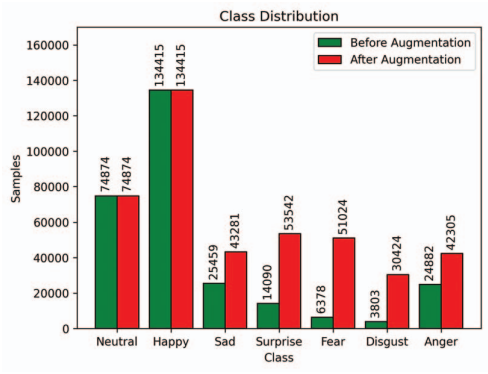
2) Data Preprocessing and Augmentation: All the training and validation samples of the dataset are of the shape 224 by 224 and are converted into RGB format. The training set of the Affectnet dataset is highly imbalanced. As shown in the plot in Fig. 2a, the number of samples in the class sad, surprise, fear, disgust, and anger are very less as compared to that in the class happy and neutral. Hence, to increase the number of samples for these classes (minority classes), we performed some image augmentations. For the two classes fear and disgust, we augmented all of the samples. For sad, surprised, and anger classes, we augmented only a fraction of the data and kept the remaining data as they were. Below listed are the transformations that we used for data augmentation.

- **Horizontal Flip:** This transformation flips our image horizontally and generates a mirrored image.
- **Gaussian Noise:** It is a kind of statistical noise that can be added to the images to reflect the images in the real world. Noisy image pixels are integrated by adding the original pixel values to a random Gaussian noise value [27].
- **Color Jitter:** This is applied to randomly change the brightness, contrast, saturation, and hue of an image.
- **Random Perspective:** This transformation randomly distorts the image by a certain degree. This augmentation is done to perceive the training image from a different perspective.
- **Aug Mix:** It mixes multiple augmentations to the image to generate diverse transformations [28].

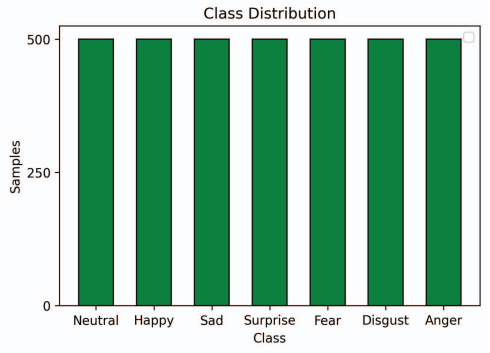
As shown in Fig. 2a, the number of data samples for class fear and disgust have increased significantly by a factor of around 10 and for classes sad, surprise, and anger by around a factor of 2 whereas for happy and neutral the number of data samples are same. This is to balance the data to some extent. However, the validation set of the AffectNet dataset is perfectly balanced as each of the seven classes has 500 images each as shown in Fig. 2b.

B. Model

1) Vision Transformer: The effective use of self-attention-based architectures like Transformers [20] in the field of



(a) Training Set



(b) Validation Set

Fig. 2: Class Distribution of AffectNet Dataset

NLP, inspired the researchers to experiment with the same architectures on computer vision as well. This led to the evolution of a transformer-based architecture in computer vision known as vision transformers. A standard transformer model in NLP applications takes in the sequence of word tokens as an input. The vision transformer implements the same standard transformer model where an image is split into a fixed-size patch of size 16 * 16 and formed as a sequence of linear embedding to feed to a transformer encoder. Hence, each image patch in the vision transformer corresponds to a word token in the NLP transformer. An extra learnable classification token is added that performs the final classification task.

2) Fine-tuning Vision Transformer: For this study, we are using a pre-trained version of the vision transformer model available publicly in the Hugging Face library. The library allows us to download and load a model *google/vit-base-patch16-224-in21k* [29], that is pre-trained on ImageNet-21k dataset [30]. For fine-tuning, we define a model that places a linear layer on top of the pre-trained model. On the top of the last hidden state of the Classifier token i.e. CLS token, it places a linear layer that can generate a representation of an input image. Also, we need to modify the output layer to match the number of outputs required for our classification task. Since, we are working with 7 emotion classes from the AffectNet Dataset (Neutral, Happy, Sad, Surprise, Fear, Disgust, and

Anger), we specified the number of output neurons of the pre-trained model to 7. The input data are preprocessed and converted into tensors using the AutoImageProcessor provided by the same Hugging Face interface. Finally, we fine-tuned the model under different settings and obtain different results.

IV. Results and Discussion

In this study, we fine-tuned the baseline ViT model under different settings by varying the number of epochs, using different optimizers and learning rates, increasing and decreasing the number of training sets, and analyzing the performance. Here, we discuss some of the major settings we performed our experiments on. However, we used the same base configuration as shown in Table I, and for training, we used NVIDIA A100-SXM4-40GB GPU.

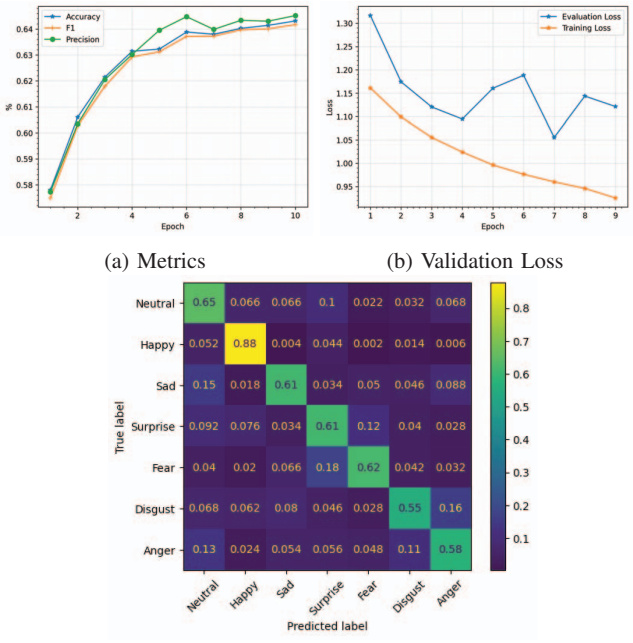
TABLE I: Base Configuration for the fine-tuning model

Model Type	Vision Transformer
Image Size	224*224*3
Hidden Size	768
Output Labels	7
Attention Heads	12
Patch Size	16*16

Initially, we trained the model with the Adam optimizer [31] with a learning rate of 0.0001 without any weight decay, for 10 epochs. We used all the training data from the AffectNet dataset without increasing or decreasing the training images and used cross-entropy loss. The result obtained when training the model or 10 epochs is shown in Fig.3. As shown in Fig 3a, we can see that the accuracy, F1 score, and precision of our model on the AffectNet validation set are converging and we achieved the best accuracy of 64.31% after the 10th epoch. Also, Fig. 3b on the right shows the converging training loss as well as the decreasing validation loss. Also, the confusion matrix in Fig. 3c, shows that the model is performing well for class happy with an accuracy of 88% but performs poorly for other classes. Mainly, anger and disgust with accuracy below 60%.

As the training and validation loss was decreasing and the accuracy was increasing, we analyzed that the model might be underperforming as we trained it for only 10 epochs. So, we increased the number of epochs to 25 as well as applied the weighted loss to prevent the model from over-fitting, and then trained our model again. For this experiment, the results are shown in Fig.4

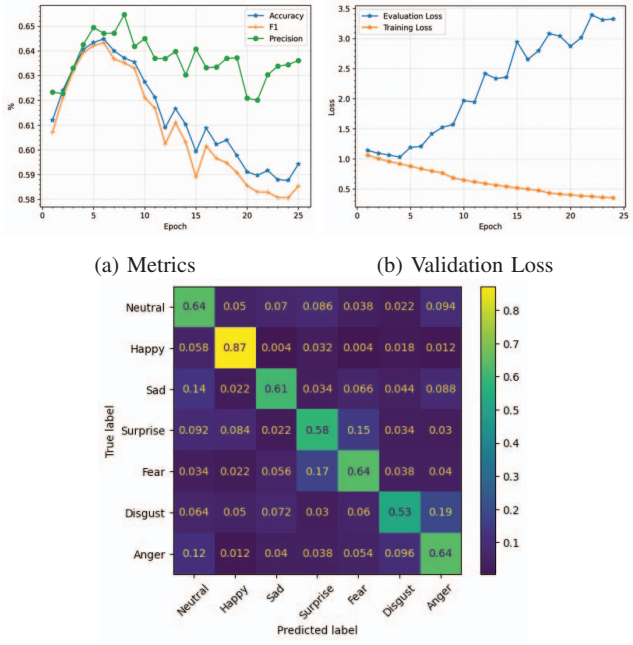
When training 25 epochs, the model achieved the best accuracy of 64.48% on the 6th epoch. After that, training loss is decreasing but validation loss seems to be increasing which is a clear sign of model over-fitting (Fig.4a and 3b). When analyzing the confusion matrix, we can see only the accuracy for class anger has improved whereas, for others, it has in fact decreased. For the class happy, the accuracy is high as before. This is because the training set has a large number of samples in the class happy as compared to other classes. Due to this, our model is not generalizing well. So, to address this issue, we performed some data augmentation to increase the number of



(a) Metrics (b) Validation Loss

(c) Confusion Matrix

Fig. 3: Performance of model when trained for 10 epochs with Adam optimizer and the learning rate of 0.0001 on original(non-augmented) data

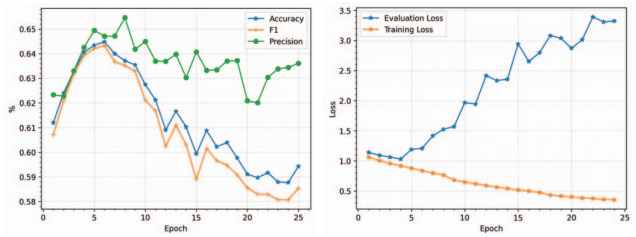


(a) Metrics (b) Validation Loss

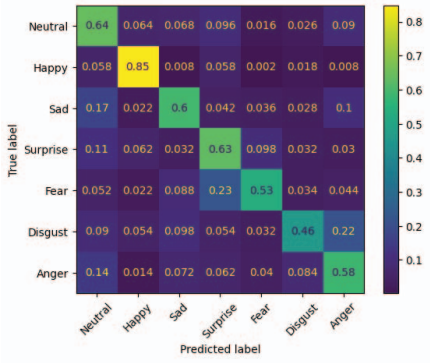
(c) Confusion Matrix

Fig. 4: Performance of model trained for 25 epochs with Adam optimizer and the learning rate of 0.0001 on original(non-augmented) data

samples for different classes by different factors whose result is shown by the red bars in Fig. 2a. The number of training data increased from 283901 to 759505 after augmentation and the training time also increased. So, we performed fine-tuning with the augmented data for 25 epochs. Also, we used Adafactor optimizer, which internally adjusts the learning rate depending on the factors like relative steps, warm-up steps, scale parameters, etc [32], in combination with the cosine scheduler with a warm-up. This scheduler adjusts the learning rate following the values of the cosine function between the initial learning set in the optimizer to 0 and after the warm-up period increases the learning rate linearly between 0 and the initially set learning rate. As we can analyze from Fig. 5, the model didn't perform any better with the augmented dataset. Although we increased the number of training samples for the images on minority classes, the accuracy for these classes is still poor. We can notice that the accuracy for class disgust has further decreased to 46%. This suggests that the augmentation technique that we implemented didn't help the model to generalize well. Our model gave an accuracy score of 61.28% only.



(a) Metrics (b) Validation Loss



(c) Confusion Matrix

Fig. 5: Performance of model trained for 25 epochs with Adafactor optimizer with an initial learning rate of 0.0001 on Augmented data

The table in II shows the performance of other transformer-based approaches and our approach on the validation set of the AffectNet database. The listed model uses Vision Transformer for FER task using a hybrid dataset. In our study, we used the entire training set of the Affectnet. However, using data augmentation to balance the training dataset did not help our model to generalize well.

TABLE II: Comparison with other transformer-based methods on AffectNet dataset. Our result is in italics.

Model	Accuracy
TFE [21]	63.33%
ViTFER [22]	54.89%
VTFF [33]	61.85 %
Ours	*64.48%*

V. Challenges and Future Work

One of the major challenges when working with transformer-based models is that they need a large dataset to generalize well and avoid over-fitting [25]. When trained on small datasets, the model tends to perform poorly even than a simple lightweight CNN model. Although the selected dataset for this research has a large number of samples, the dataset is quite unbalanced. The number of samples for the happy class (∼135k) and neutral class(∼75k) is very large as compared to that of the class disgust (∼4k) and fear class(∼6k). Moreover, the augmentation technique we used did not work well. So, to address these issues, we plan to study on more effective augmentation technique to balance the dataset.

The main purpose of this research is to study facial emotion recognition in the wild so that it can be integrated into the vehicular system to improve driver safety and comfort. Many studies like [34], [35] etc. show that the emotions like anger, nervousness, sadness and being hostile are directly associated with aggressive driving. Negative emotions can hamper the cognitive process and risk the safety of drivers and the environment around them. Hence, we can use the proper FER system to classify the different moods of the drivers and act accordingly. Different approaches can be taken then such as playing the music as per the mood, alarming the driver of their speed or irregular driving, and so on. Also, in the future, we need to train a model on the domain-specific dataset such as Driver Emotion Facial Expression Dataset [36]. This will make the system more effective. Further, we are currently fine-tuning a model that is pre-trained on the Imagenet-21k dataset which has a large number of classes. We plan on pre-training our own model using the face recognition dataset so that the pre-trained model can have already learned facial features. Then this model can be fine-tuned for FER-specific tasks. Finally, we aim to focus on handling the occlusions that are widely seen in the real world.

VI. Conclusion

In this study, we fine-tuned the Vision Transformer's baseline model initially trained on Imagenet21k on the AffectNet dataset to classify the emotion of a person from facial expression. The dataset used in this study is one of the largest publicly available datasets for FER tasks with more than 1 million images. Due to the variation in the number of samples per class, we augmented the images of the classes

with lower samples to increase their number. However, we achieved the best accuracy of 64.48% when fine-tuning on a non-augmented dataset. This study is the preliminary study to exploit the use of facial Emotion recognition tasks in human-machine teaming applications. One such application can be in an intelligent vehicular system where different actions can be triggered based on the driver's emotions such as recommending music. This can help to improve the driver's safety as well as comfort to enhance their experiences. Hence, using the approach discussed in this research, a more effective and accurate FER system can be developed.

REFERENCES

[1] D. H. Hagos and D. B. Rawat, "Recent advances in artificial intelligence and tactical autonomy: Current status, challenges, and perspectives," *Sensors*, vol. 22, no. 24, p. 9916, 2022.

[2] D. B. Rawat, "Artificial intelligence meets tactical autonomy: Challenges and perspectives," in *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2022, pp. 49–51.

[3] H. El Alami, M. Nwosu, and D. B. Rawat, "Joint human and autonomy teaming for defense: status, challenges, and perspectives," *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, vol. 12538, pp. 144–158, 2023.

[4] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *arXiv*, Apr. 2018.

[5] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional Expression: Advances in Basic Emotion Theory," *Journal of nonverbal behavior*, vol. 43, no. 2, p. 133, Jun. 2019.

[6] P. Ekman and W. V. Friesen, "Facial Action Coding System," Jan. 2019, institution: American Psychological Association. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/t27734-000

[7] F. Abdat, C. Maaoui, and A. Pruski, "Human-Computer Interaction Using Emotion Recognition from Facial Expression," in *2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, Nov. 2011, pp. 196–201.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2021.

[9] 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. PP, no. 99, pp. 1–1, 2017.

[10] M. Nazir, Z. Jan, and M. Sajjad, "Facial expression recognition using histogram of oriented gradients based transformed features," *Cluster Computing*, vol. 21, no. 1, pp. 539–548, Mar. 2018.

[11] D. Ghimire, J. Lee, Z.-N. Li, and S. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7921–7946, Mar. 2017.

[12] Z. Song, "Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory," *Frontiers in Psychology*, vol. 12, Sep. 2021.

[13] V. Ch, U. S. Reddy, and V. K. K. Kolli, "Facial emotion recognition using nlpca and svm," *Traitement du Signal*, vol. 36, pp. 13–22, 04 2019.

[14] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, p. 268, May 2022. [Online]. Available: http://dx.doi.org/10.3390/info13060268

[15] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[16] P. Antoniadis, P. P. Filntisis, and P. Maragos, "Exploiting emotional dependencies with graph convolutional networks for facial expression recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021, pp. 1–8.

[17] E. Ryumina, D. Dresvyanskiy, and A. Karpov, "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study," *Neurocomputing*, vol. 514, pp. 435–450, Dec. 2022.

[18] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 67–74.

[19] 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, 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[21] J. Gao and Y. Zhao, "TFE: A Transformer Architecture for Occlusion Aware Facial Expression Recognition," *Frontiers in Neurorobotics*, vol. 15, Oct. 2021.

[22] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "Vitfer: Facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, Aug 2022. [Online]. Available: http://dx.doi.org/10.3390/asi5040080

[23] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," 2013.

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

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[26] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=4nPswr1KcP

[27] V. Lendave, "A Guide to Different Types of Noises and Image Denoising Methods," *Analytics India Magazine*, Sep. 2021. [Online]. Available: https://analyticsindiamag.com/a-guide-to-different-types-of-noises-and-image-denoising-methods

[28] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," 2020.

[29] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[32] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," 2018.

[33] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[34] N. Kovácsová, T. Lajunen, and E. Rošková, "Aggression on the road: Relationships between dysfunctional impulsivity, forgiveness, negative emotions, and aggressive driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 42, pp. 286–298, Oct. 2016.

[35] T. Zimasa, S. Jamson, and B. Henson, "Are happy drivers safer drivers? Evidence from hazard response times and eye tracking data," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 46, pp. 14–23, Apr. 2017.

[36] W. Li, Y. Cui, Y. Ma, X. Chen, G. Li, G. Zeng, G. Guo, and D. Cao, "A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 747–760, 2023.