

FE-Adapter: Adapting Image-based Emotion Classifiers to Videos

Shreyank N Gowda¹, Boyan Gao¹ and David A. Clifton^{1,2}

¹ Department of Engineering Sciences, University of Oxford, OX3 7DQ Oxford, UK

² Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou 215123, Jiangsu, China

Abstract—Utilizing large pre-trained models for specific tasks has yielded impressive results. However, fully fine-tuning these increasingly large models is becoming prohibitively resource-intensive. This has led to a focus on more parameter-efficient transfer learning, primarily within the same modality. But this approach has limitations, particularly in video understanding where suitable pre-trained models are less common. Addressing this, our study introduces a novel cross-modality transfer learning approach from images to videos, which we call parameter-efficient image-to-video transfer learning. We present the Facial-Emotion Adapter (FE-Adapter), designed for efficient fine-tuning in video tasks. This adapter allows pre-trained image models, which traditionally lack temporal processing capabilities, to analyze dynamic video content efficiently. Notably, it uses about 15 times fewer parameters than previous methods, while improving accuracy. Our experiments in video emotion recognition demonstrate that the FE-Adapter can match or even surpass existing fine-tuning and video emotion models in both performance and efficiency. This breakthrough highlights the potential for cross-modality approaches in enhancing the capabilities of AI models, particularly in fields like video emotion analysis where the demand for efficiency and accuracy is constantly rising.

I. INTRODUCTION

Deep learning has risen to prominence as a transformative technology, delivering groundbreaking achievements in multiple fields. It has revolutionized tasks ranging from image recognition [15], [12], [7], emotion recognition [28], [1], [6] and action recognition [11], [10], [2], pushing the limits of machine capabilities. These significant improvements in precision have altered the landscape of numerous industries, providing innovative solutions to problems that have persisted for years.

The rapid evolution of self-supervised models in machine learning, particularly in the domains of image [33], [31] and language processing, has marked a significant advancement in the field. These models have demonstrated remarkable proficiency in a wide range of downstream tasks, setting new benchmarks across various applications. However, as these models grow in size, the traditional method of fully fine-tuning them for specific tasks becomes increasingly untenable [16]. This is predominantly due to the exorbitant computational costs and storage requirements associated with such processes.

Recognizing this challenge, the focus of recent research has shifted towards more efficient methods of transfer learning. The objective is to leverage the knowledge captured in these pre-trained models while minimizing the additional resources required for adaptation to new tasks [16], [30]. Existing strategies predominantly concentrate on tasks within

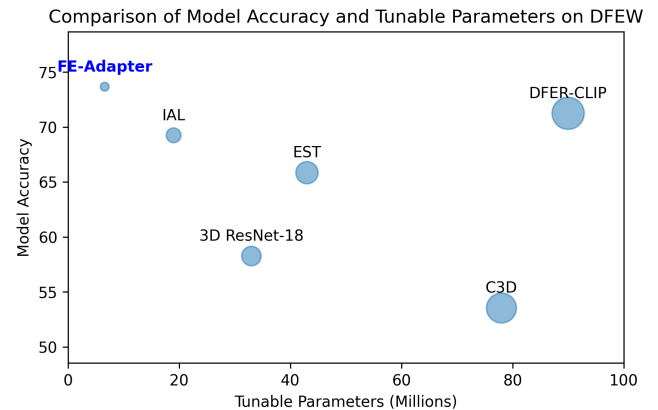


Fig. 1. A comparative analysis of various video-based models on the DFEW [18] dataset, showcasing the correlation between the number of tunable parameters (in millions) and model accuracy (%). Our proposed FE-Adapter (highlighted in blue and bold), requires significantly fewer trainable parameters whilst outperforming recent SOTA models including vision-language models. The size of each bubble represents the number of tuneable parameters of the respective model. We compare with recent SOTA models such as IAL [22], EST [27], and DFER-CLIP [38]. We also compare against older 3D based models such as 3D-ResNet [15] and C3D [34].

the same modality as the pre-trained model, such as image processing [26], [38]. This approach, while effective, has inherent limitations, especially in modalities like video understanding where equivalent pre-trained models are either scarce or non-existent.

The introduction of vision transformers (ViTs) [7] has resulted in significant improvements in general computer vision problems. However, ViTs often perform inferiorly compared to CNNs due to a lack of inductive bias and a tendency to focus on occlusions and noisy areas [1]. Specific approaches to improving ViTs for facial emotion recognition includes using attentional selective fusion [28], squeeze and excitation blocks [1] or neural resizing [17] among other approaches. Nonetheless, the application of ViTs in videos is very scarce.

In this context, our study introduces a novel concept in cross-modality transfer learning, explicitly transitioning from facial-image emotion understanding to video emotion understanding. We present the Facial-emotion Adapter (FE-Adapter), a bespoke solution designed for efficient fine-tuning in video-related tasks. The FE-Adapter stands out with its compact and effective architecture. It enables pre-trained image models, which inherently lack the capability to process temporal dynamics, to interpret and analyze video content effectively. Figure 1 highlights the significant difference

in tuneable parameters (size of bubble and x-axis) whilst improving upon the state-of-the-art.

One of the most compelling aspects of the FE-Adapter is its parameter efficiency. It requires approximately 8% of the parameters per task (in comparison to current state-of-the-art models), translating to about 15 times fewer updated parameters than the state-of-the-art [38]. This feature addresses the critical challenge of resource-intensive fine-tuning in large models. Through our extensive experiments in the realm of video emotion recognition, the FE-Adapter has demonstrated its ability to not only match but, in some instances, surpass both the traditional comprehensive fine-tuning approaches and the latest models dedicated to video emotion. This achievement is particularly noteworthy as it maintains the dual benefits of parameter efficiency and adaptability, making the FE-Adapter an invaluable tool in the field of video emotion recognition, where efficiency and flexibility are paramount.

II. RELATED WORK

A. Emotion Recognition in Image

Emotion recognition in images has progressed significantly with the advent of deep learning. Convolutional Neural Networks (CNNs) have been extensively utilized [8], [19], [21] for this task due to their ability to capture spatial hierarchies in facial features.

Image classification models have been adapted for facial emotion recognition (FER), but challenges persist [6] in natural environments due to issues like pose variations, occlusions, and distracting backgrounds. To tackle these, attention-based and region-based methods have emerged. Attention-based techniques focus on the most informative parts of facial expressions, while region-based strategies segment the face into sub-regions using various cropping techniques to minimize occlusions and background interference.

However, they often overlook the interconnectivity between these regions. Vision Transformers (ViTs) [7] have brought notable advancements in computer vision, but they sometimes fall short in comparison to CNNs, especially in handling occlusions and noisy areas. Methods to enhance ViTs for FER involve attentional selective fusion [28], squeeze and excitation blocks [1], and neural resizing [17]. Yet, their application in video-based FER remains limited.

B. Emotion Recognition in Videos

Video emotion recognition (VER) analyzes emotions in videos by combining temporal and spatial data. It differs from image recognition, using 3D CNNs [34], [13] to track spatial and temporal variations in facial expressions.

VER faces challenges like pose changes, lighting, and obstructions, exacerbated by motion blur and varying frame rates. Solutions include a convolutional spatial transformer and a temporal transformer [37], handling occlusions and varied poses. Other approaches process videos in short clips [26], employing clip-based encoders and emotional

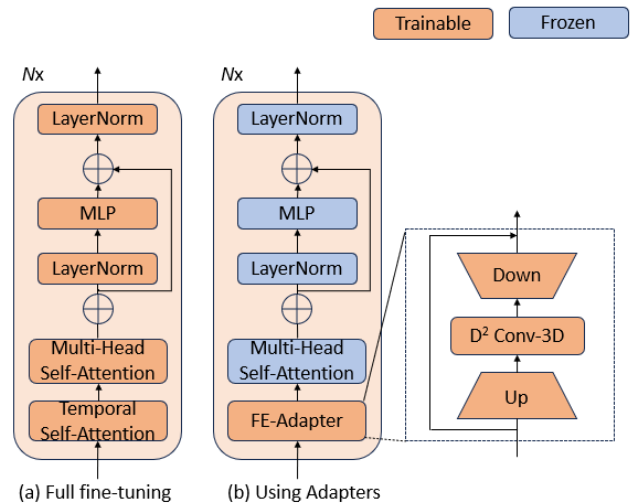


Fig. 2. Adapters ensure minimal parameter updates whilst keeping the generalization ability of the pre-trained model consistent. In comparison to (a) full fine-tuning, using (b) adapter modules significantly reduces tuneable parameter count.

intensity networks for emotion detection, or using intensity-aware losses [22] and Multi-Instance Learning [35] to manage imprecise labels and capture time-based relationships.

Self-supervised pre-training [33] on large-scale data with a Transformer encoder is also used. DFER-CLIP [38], a visual-language model based on CLIP [31], specifically focuses on Dynamic Facial Expression Recognition (DFER), using a Transformer and textual descriptions to improve accuracy. While CLIP has been extensively adapted for video action recognition, its application in VER is less explored.

C. Adapting Image models to Video

The growing use of large pre-trained language models for various tasks has highlighted the importance of efficient tuning in NLP. This form of parameter-efficient tuning can be broadly categorized into two main approaches.

First is the use of task-specific adapters [16], [14], which are light modules added between the layers of a pre-trained model. For efficiency, only these adapters are updated during task fine-tuning, leaving the bulk of the pre-trained model's parameters unchanged.

The second approach is prompt tuning [24], [20], where learnable tokens are added either at the model's input or intermediate layers, and only these tokens are adjusted for each task whilst obtaining impressive results.

While approaches have been proposed [30], [9] to adapt image models to videos for tasks in video understanding, to the best of our knowledge nothing has been proposed for video emotion recognition.

III. METHODOLOGY

A. PRELIMINARIES

To adapt highly-trained image models for video processing, it is essential to bridge the gap between still images and videos. We explore methods to aggregate image-based

features for videos, with a focus on adapting Vision Transformers (ViTs) to video models like Timesformer [4].

In the simplest scenario, a pre-trained image model for video understanding involves aggregating frame features over time, such as through average pooling. Consider a video clip V defined as $V \in \mathbb{R}^{T \times H \times W}$, where T , H , and W represent the number of frames, height, and width respectively. Each frame is divided into $N = \frac{H \times W}{P^2}$ patches of size $P \times P$. These patches are flattened and projected into patch tokens $Z_t = [z_1, \dots, z_s, \dots, z_N]$, with $z_s \in \mathbb{R}^d$ and $d = 3 \times P^2$, for $t = 1, \dots, T$. Positional embeddings are added to this series of feature vectors, along with a trainable class token. Each sequence of $N+1$ tokens undergoes self-attention processing, retaining only the classification token z_{cls}^t for each sequence. Temporal average pooling is then applied to these class tokens $z_{\text{final}} = \frac{1}{T} \sum_t z_{\text{cls}}^t$, to form a condensed representation for the entire clip. The final prediction is derived from a classifier using z_{final} . This method, known as the Space-Only TimeSformer [4], averages spatial information over time.

Applying ViTs to the video domain involves enhancing spatio-temporal attention with additional temporal layers [2], [4], [5]. While these models achieve higher accuracy, they typically require full fine-tuning for each task. This approach is inefficient in terms of parameters as it demands maintaining separate fully fine-tuned model parameters for every task.

B. ADAPTERS

Adapter modules, initially prominent in natural language processing (NLP), have been increasingly integrated into computer vision applications [16]. These modules are particularly known for their parameter efficiency, which means they add a relatively small number of trainable parameters to a pre-existing model to tailor it for specific tasks. This approach allows for adapting models to new tasks or datasets without the need to retrain the entire model, thus saving computational resources and time.

In the context of computer vision, adapter modules are inserted into pre-trained neural networks. These networks, designed for general tasks, are fine-tuned to perform specific vision tasks such as object detection, image segmentation, or image classification. The core idea behind adapters is to make these networks more flexible and efficient in learning task-specific features.

Given an input feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ at the i -th layer of a neural network, the adaptation process performed by the adapter module can be formally defined as:

$$\text{Adapter}(\mathbf{X}) = \mathbf{X} + f(\mathbf{X}W_{\text{down}})W_{\text{up}}, \quad (1)$$

where $W_{\text{down}} \in \mathbb{R}^{d \times r}$ denotes the down-projection layer, $W_{\text{up}} \in \mathbb{R}^{r \times d}$ the up-projection layer, and $f(\cdot)$ represents the activation function. The down-projection layer reduces the dimensionality of the input features \mathbf{X} to a lower-dimensional space \mathbb{R}^r , aiming to capture the most relevant information for the task at hand. The up-projection layer then projects these features back to the original high-dimensional space \mathbb{R}^d . This process is designed to enrich the original features with task-specific information. A residual connection

adds the output of the adapter module to the original input features \mathbf{X} , ensuring that the information present in the initial features is not lost, thereby conserving the input information.

The features \mathbf{X} referred to in this context are typically the embeddings or representations extracted from the i -th layer of the neural network. These embeddings encode the input data (e.g., images) into a form that the network can process to perform tasks like classification or detection. By applying adapter modules to these embeddings, the network can learn to adjust its representations to better suit specific tasks or datasets, enhancing performance with minimal additional trainable parameters.

C. FE-ADAPTER

The Adapter module, notably the ST-Adapter [30], has been adapted for image models in video tasks. However, adapting facial image emotion models to videos remains unexplored. We propose using adapters as illustrated in Figure 2.

ST-Adapter uses depth-wise 3D-convolution for spatio-temporal analysis. This method might not be ideal for video emotion recognition due to its limitations in capturing complex spatial and temporal interactions, as it processes each channel independently, potentially overlooking emotional cues in facial expressions and body language. Depth-wise convolutions might not integrate spatial and temporal patterns effectively, which are crucial for emotion recognition.

We propose Dynamic Dilated Convolutions (D²Conv3D) [32] for enhanced performance in video emotion recognition tasks. The dynamic dilation rates of D²Conv3D are well-suited for capturing intricate spatio-temporal patterns, especially in subtle facial emotions and temporal dynamics. This makes it more suitable for accurate emotion detection. The enhanced feature extraction capabilities of D²Conv3D make it ideal for analyzing nuanced emotional cues in videos. We define our FE-Adapter as:

$$\text{FE-Adapter}(X) = X + f(D^2\text{Conv3D}(XW_{\text{down}}))W_{\text{up}} \quad (2)$$

Our Ablation Study highlights the importance of this choice for emotion recognition. In integrating the FE-Adapter for video emotion recognition, similar to adapters in NLP placed between Transformer layers [16], various strategies are considered. Typically in NLP, adapters are positioned after both the Multi-Head Self-Attention (MHSA) and the Multi-Layer Perceptron (MLP). Some studies [16], [30] suggest that a single adapter post-MLP is effective.

For the FE-Adapter, we explore similar placements. Empirical evidence suggests that positioning a single FE-Adapter before the MHSA in each transformer block can enhance performance, allowing the FE-Adapter to preprocess input features effectively before the MHSA, and optimizing emotion recognition.

IV. EXPERIMENTAL ANALYSIS

A. DATASETS

For our video emotion recognition task, we utilize a range of datasets, each offering unique characteristics and

Dataset	DFEW [18]		FERV39k [36]		MAFW [25]	
	UAR	WAR	UAR	WAR	UAR	WAR
RAF-DB	60.59	72.31	40.04	50.92	38.15	52.25
FERPlus	60.66	72.42	40.44	51.31	38.87	53.48
AffectNet	60.89	73.67	41.44	52.11	39.41	55.02

TABLE I

EFFECT OF IMAGE MODELS FINE-TUNED ON DIFFERENT IMAGE DATASETS.

Method	DFEW [18]		FERV39k [36]		MAFW [25]	
	UAR	WAR	UAR	WAR	UAR	WAR
TA	53.66	68.51	31.92	46.41	31.14	47.51
Linear Probe	53.85	69.66	33.12	47.22	34.59	50.17
DW-Conv3D	57.65	71.15	37.69	50.40	37.11	51.91
D ² Conv3D	60.89	73.67	41.44	52.11	39.41	55.02

TABLE II

COMPARING TEMPORAL CONVOLUTION CHOICES. ‘TA’ REFERS TO TEMPORAL AGGREGATION.

challenges. These include DFEW [18], FERV39K [36], and MAFW [25]. DFEW features 16,000 video clips across seven emotions, with 12,059 clips used for 5-fold cross-validation experiments. FERV39K provides 38,935 video sequences derived from four different scenarios, split into 31,088 for training and 7,847 for testing. MAFW, focusing on the video modality, comprises 10,045 video clips annotated with 11 emotions, using 9,172 clips in a 5-fold cross-validation setup. These diverse datasets allow for comprehensive training and evaluation of our emotion recognition models and also is commonly used in current state-of-the-art models.

B. IMPLEMENTATION DETAILS

In our experiments, we employ PF-ViT based on the ViT-B architecture, re-implemented with identical hyperparameters for AffectNet classification. We adapt this model for video tasks, using one adapter per PF-ViT block. The training involves the AdamW optimizer with a learning rate of 5e-4 and weight decay of 1e-2, alongside a batch size of 128. We process 16 frames per video, resized to 224×224. The learning rate is decayed using cosine annealing over 100 epochs. All experiments are conducted using an NVIDIA A100 GPU.

C. ABLATION STUDY

1) *Choice of Pre-Training Dataset:* We use PF-ViT fine-tuned on AffectNet [29] for adapting to video datasets. We also compare the use of RAF-DB [23] and FERPLUS [3]. These results are shown in Table I.

2) *Comparing Temporal Convolution Choices:* In comparison to ST-Adapter, we use a D²Conv3D instead of DW-Conv3D. Based on this, we compare the performance of different choices including a simple temporal aggregation of frame-by-frame features. We see a significant improvement using D²Conv3D proving our hypothesis that discerns subtle facial emotions and temporal emotional dynamics. These results are in Table II.

1-4	5-8	9-12	FERV39k [36]		MAFW [25]	
			UAR	WAR	UAR	WAR
✓			35.41	45.67	34.15	48.22
	✓		38.85	49.81	36.69	51.12
		✓	40.44	51.08	38.22	53.65
	✓	✓	40.89	51.66	39.00	54.11
✓	✓	✓	41.44	52.11	39.41	55.02

TABLE III

COMPARING GLOBAL ADAPTER POSITION CHOICES.

Position	FERV39k [36]		MAFW [25]	
	UAR	WAR	UAR	WAR
After MLP	41.35	51.89	39.25	54.78
After MHSA	41.39	52.07	39.33	54.92
Before MHSA	41.44	52.11	39.41	55.02

TABLE IV

COMPARING LOCAL ADAPTER POSITION CHOICES.

3) *Adapter Position:* We consider two scenarios in terms of where we position the adapter. We first consider the global position i.e. what blocks do we place the adapter in. We do this by breaking the ViT-B to three sub-blocks such as 1-4, 5-8 and 9-12. We see that using an adapter in every block gives us best results. These results are in Table III.

We also consider where to place it within each block. With this we have a choice of placing it before the Multi-Head Self-Attention (MHSA) or after. We could also place it after the MLP. These results are shown in Table IV.

D. COMPARISON TO STATE-OF-THE-ART

We closely examine the performance of our proposed FE-Adapter against various leading models in the domain of facial emotion recognition. Table V presents a comparative analysis of various methods in the context of facial emotion recognition using three popular benchmarks: DFEW, FERV39k, and MAFW. Each method is evaluated based on its performance in Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) across these datasets. The table also includes the number of tunable parameters (in millions) for each method, indicating the complexity and potential computational demands.

Notable methods listed include C3D [34], 3D ResNet-18 [13], Former-DFER [37], CEFLNet [26], IAL [22], EST [27], M3DFEL [35], MAE-DFER [33] and DFER-CLIP [38]. A key observation is the relatively low number of tunable parameters in the FE-Adapter method (6.6 million), which is significantly lower than most other methods.

Overall, the table illustrates that FE-Adapter, despite its markedly lower complexity, competes well with or outperforms recent state-of-the-art models in facial emotion recognition across different benchmarks, highlighting its effectiveness and efficiency.

V. ACKNOWLEDGEMENTS

DAC was supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research

Method	Tunable Params (M)	DFEW [18]		FERV39k [36]		MAFW [25]	
		UAR	WAR	UAR	WAR	UAR	WAR
C3D [34] [ICCV'15]	78	42.74	53.54	22.68	31.69	31.17	42.25
3D ResNet-18 [13] [ICCVW'17]	33	46.52	58.27	26.67	37.57	-	-
Former-DFER [37] [MM'21]	18	53.69	65.70	37.20	46.85	-	-
CEFLNet [26] [IS'2022]	13	51.14	65.35	-	-	-	-
IAL [22] [AAAI'23]	19	55.71	69.24	35.82	48.54	-	-
EST [27] [PR'23]	43	53.43	65.85	-	-	-	-
M3DFEL [35] [CVPR'23]	-	56.10	69.25	35.94	47.67	-	-
MAE-DFER [33] [MM'23]	85	63.41	74.43	43.12	52.07	41.62	54.31
DFER-CLIP [38] [BMVC'23]	90	59.61	71.25	41.27	51.65	38.89	52.55
FE-Adapter (Ours)	6.6	60.89	73.67	41.44	52.11	39.41	55.02

TABLE V

COMPARING FE-ADAPTER WITH STATE-OF-THE-ART MODELS USING UAR (UNWEIGHTED AVERAGE RECALL) AND WAR (WEIGHTED AVERAGE RECALL) ON THREE POPULAR BENCHMARKS.

Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Wellcome Trust funded VITAL project (grant 204904/Z/16/Z); the EPSRC (grant EP/W031744/1); and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE).

VI. CONCLUSION

In conclusion, this study presents an innovative cross-modality transfer learning approach, namely parameter-efficient image-to-video transfer learning, through the development of the Facial-Emotion Adapter (FE-Adapter). This novel approach addresses the challenges posed by the conventional method of fully fine-tuning large pre-trained models, which is resource-intensive in terms of training and storage. The FE-Adapter demonstrates the feasibility and effectiveness of using pre-trained image models for video understanding tasks. It does so by equipping these models with the capability to process dynamic video content, while significantly reducing the requirement for updated parameters – approximately 15 times fewer than previous methods. Our comprehensive experiments in video emotion recognition indicate that the FE-Adapter not only holds its own against but in some cases, outperforms both the traditional comprehensive fine-tuning methods and the latest models in this domain. Importantly, it achieves these results while upholding the advantages of parameter efficiency. This makes the FE-Adapter a highly valuable contribution to the field of video emotion recognition, offering a practical solution that combines efficiency with adaptability.

REFERENCES

- [1] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Segurier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021.
- [2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.
- [3] 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*, pages 279–283, 2016.
- [4] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
- [5] A. Bulat, J. M. Perez Rua, S. Sudhakaran, B. Martinez, and G. Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in neural information processing systems*, 34:19594–19607, 2021.
- [6] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [8] S. N. Gowda. Face verification across age progression using facial feature extraction. In *2016 International Conference on Signal and Information Processing (ICONSIP)*, pages 1–5. IEEE, 2016.
- [9] S. N. Gowda, A. Arnab, and J. Huang. Optimizing vivit training: Time and memory reduction for action recognition. *arXiv preprint arXiv:2306.04822*, 2023.
- [10] S. N. Gowda, M. Rohrbach, F. Keller, and L. Sevilla-Lara. Learn2augment: Learning to composite videos for data augmentation in action recognition. In *European conference on computer vision*, pages 242–259. Springer, 2022.
- [11] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara. Smart frame selection for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1451–1459, 2021.
- [12] S. N. Gowda and C. Yuan. Colornet: Investigating the importance of color spaces for image classification. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 581–596. Springer, 2019.
- [13] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.
- [14] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.
- [15] 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*, pages 770–778, 2016.
- [16] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [17] H. Hwang, S. Kim, W.-J. Park, J. Seo, K. Ko, and H. Yeo. Vision transformer equipped with neural resizer on facial expression recognition task. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2614–2618. IEEE, 2022.
- [18] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.
- [19] B. C. Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.
- [20] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*,

- 2021.
- [21] B. Li and D. Lima. Facial expression recognition via resnet-50. *International Journal of Cognitive Computing in Engineering*, 2:57–64, 2021.
 - [22] H. Li, H. Niu, Z. Zhu, and F. Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 67–75, 2023.
 - [23] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.
 - [24] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.
 - [25] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 24–32, 2022.
 - [26] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, and Z. Luo. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences*, 598:182–195, 2022.
 - [27] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, and Y. Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023.
 - [28] F. Ma, B. Sun, and S. Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021.
 - [29] 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*, 10(1):18–31, 2017.
 - [30] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.
 - [31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
 - [32] C. Schmidt, A. Athar, S. Mahadevan, and B. Leibe. D2conv3d: Dynamic dilated convolutions for object segmentation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1200–1209, 2022.
 - [33] L. Sun, Z. Lian, B. Liu, and J. Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023.
 - [34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
 - [35] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17958–17968, 2023.
 - [36] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.
 - [37] Z. Zhao and Q. Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021.
 - [38] Z. Zhao and I. Patras. Prompting visual-language models for dynamic facial expression recognition. *The British Machine Vision Conference*, 2023.