Rethinking the Learning Paradigm for Dynamic Facial Expression Recognition

Hanyang Wang[12*], Bo Li[3*†], Shuang Wu[3], Siyuan Shen[12], Feng Liu[412†], Shouhong Ding[3], Aimin Zhou[12†]
[1]Shanghai Institute of AI for Education, East China Normal University
[2]School of Computer Science and Technology, East China Normal University
[3]Youtu Lab, Tencent
[4]Shanghai International School of Chief Technology Officer, East China Normal University

Abstract

Dynamic Facial Expression Recognition (DFER) is a rapidly developing field that focuses on recognizing facial expressions in video format. Previous research has considered non-target frames as noisy frames, but we propose that it should be treated as a weakly supervised problem. We also identify the imbalance of short- and long-term temporal relationships in DFER. Therefore, we introduce the Multi-3D Dynamic Facial Expression Learning (M3DFEL) framework, which utilizes Multi-Instance Learning (MIL) to handle inexact labels. M3DFEL generates 3D-instances to model the strong short-term temporal relationship and utilizes 3DCNNs for feature extraction. The Dynamic Long-term Instance Aggregation Module (DLIAM) is then utilized to learn the long-term temporal relationships and dynamically aggregate the instances. Our experiments on DFEW and FERV39K datasets show that M3DFEL outperforms existing state-of-the-art approaches with a vanilla R3D18 backbone. The source code is available at https://github.com/faceeyes/M3DFEL.

1. Introduction

Facial expressions are essential in communication [26, 27, 45]. Understanding the emotions of others through their facial expressions is critical during conversations. Thus, automated recognition of facial expressions is a significant challenge in various fields, such as human-computer interaction (HCI) [25, 34], mental health diagnosis [12], driver fatigue monitoring [24], and metahuman [6]. While significant progress has been made in Static Facial Expression Recognition (SFER) [23, 43, 44, 55], there is increasing attention on Dynamic Facial Expression Recognition.

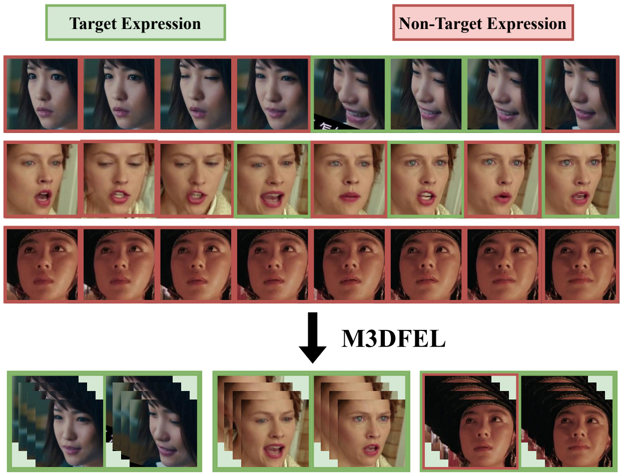


Figure 1. In-the-wild Dynamic Facial Expressions. In the first row of images, the subject appear predominantly *Neutral*, yet the video is labeled as *happy* without specifying the exact moment when the emotion is expressed. In the second row, the emotion is evident from the perspective of a few figures, but any single one of them is noisy and unclear. In the third row, all frames appear *Neutral*, but a closer analysis of facial movement over time reveals a rising of the corner of the mouth, indicating a smile.

With the availability of large-scale in-the-wild datasets like DFEW [11] and FERV39K [46], several methods have been proposed for DFER [21, 22, 31, 47, 54]. Previous works [31, 54] have simply applied general video understanding methods to recognize dynamic facial expressions. Later on, Li *et al.* [22] observe that DFER contains a large number of noisy frames and propose a dynamic class token and a snippet-based filter to suppress the impact of these frames. Li *et al.* [21] propose an Intensity Aware Loss to account for the large intra-class and small inter-class differences in DFER and force the network to pay extra attention to the most confusing class. However, we argue that DFER requires specialized designs rather than being con-

[*] Both author contributed equally to this work. Work done during Hanyang Wang's internship at Tencent Youtu Lab and Bo Li is the project lead.

[†] Corresponding authors. libraboli@tencent.com, lsttoy@163.com, amzhou@cs.ecnu.edu.cn

sidered a combination of video understanding and SFER. Although these works [21, 22, 47] have identified some issues in DFER, their models have only addressed them in a rudimentary manner.

Firstly, these works fail to recognize that the existence of non-target frames in DFER is actually caused by weak supervision. When collecting large-scale video datasets, annotating the precise location of labels is labor-intensive and challenging. A dynamic facial expression may contain a change between non-target and target emotions, as shown in Figure 1. Without a location label that can guide the model to ignore the irrelevant frames and focus on the target, models are likely to be confused by the inexact label. Therefore, modeling these non-target frames as noisy frames directly is superficial, and the underlying weakly supervised problem remains unsolved.

Secondly, the previous works directly follow to use sequence models without a dedicated design for DFER. However, we find that there is an imbalance between short- and long-term temporal relationships in DFER. For example, some micro-expressions may occur within a short clip, while some facial movements between expressions may disrupt individual frames, as shown in Figure 1. In contrast, there is little temporal relationship between a *Happy* face at the beginning of a video and another *Happy* face at the end. Therefore, neither modeling the entire temporal relationship nor using completely time-irrelevant aggregation methods is suitable for DFER. Instead, a method should learn to model the strong short-term temporal relationship and the weak long-term temporal relationship differently.

To address the first issue, we suggest using weakly supervised strategies to train DFER models instead of treating non-target frames as noisy frames. Specifically, we propose modeling DFER as a Multi-Instance Learning (MIL) problem, where each video is considered as a bag containing a set of instances. In this MIL framework, we disregard non-target emotions in a video and only focus on the target emotion. However, most existing MIL methods are time-independent, which is unsuitable for DFER. Therefore, a dedicated MIL framework for DFER is necessary to address the imbalanced short- and long-term temporal relationships.

The M3DFEL framework proposed in this paper is designed to address the imbalanced short- and long-term temporal relationships and the weakly supervised problem in DFER in a unified manner. It uses a combination of 3D-Instance and R3D18 models to enhance short-term temporal learning. Once instance features are extracted, they are fed into the Dynamic Long-term Instance Aggregation Module (DLIAM), which aggregates the features into a bag-level representation. The DLIAM is specifically designed to capture long-term temporal relationships between instances. Additionally, the Dynamic Multi-Instance Normalization (DMIN) is employed to maintain temporal consistency at both the bag-level and instance-level by performing dynamic normalization.

Overall, our contributions can be summarized as follows:

- We propose a weakly supervised approach to model Dynamic Facial Expression Recognition (DFER) as a Multi-Instance Learning (MIL) problem. We also identify an imbalance between short- and long-term temporal relationships in DFER, which makes it inappropriate to model the entire temporal relationship or use time-irrelevant methods.
- We propose the Multi-3D Dynamic Facial Expression Learning (M3DFEL) framework to provide a unified solution to the weakly supervised problem and model the imbalanced short- and long-term temporal relationships in DFER.
- We conduct extensive experiments on DFEW and FERV39K, and our proposed M3DFEL achieves state-of-the-art results compared with other methods, even when using a vanilla R3D18 backbone. We also conduct visualization experiments to analyze the performance of M3DFEL and uncover unsolved problems.

2. Related Work

2.1. Dynamic Facial Expression Recognition

Following the success of DNNs in computer vision tasks [3, 4, 15–20, 35–38, 50, 51, 56, 57], automatic Facial Expression Recognition (FER) has been improved via Deep learning. DFER methods differ from SFER methods as they need to consider temporal information in addition to spatial features in each image. Some methods employ CNNs to extract spatial features from each frame and then use RNNs to analyze the temporal relationship [28, 52]. 3DCNNs have been proposed to model 3D data and learn spatial and temporal features jointly. Fan *et al*. [5] proposed a hybrid network that combines recurrent neural networks (RNN) and 3D convolutional networks (C3D) using late fusion. Lee *et al*. [14] proposed a scene-aware hybrid neural network that combines 3DCNNs, 2DCNNs, and RNNs in a novel way. Lee *et al*. [13] presented CAER-Net, a deep network for context-aware emotion recognition that exploits both human facial expression and context information in a joint and boosting manner.

Recently, transformer-based networks have gained popularity in extracting both spatial and temporal information. For example, Zha *et al*. [54] propose a dynamic facial expression recognition transformer (Former-DFER) which consists of a convolutional spatial transformer (CS-Former) and a temporal transformer (T-Former). Ma *et al*. [31] propose the spatial-temporal Transformer (STT) to capture discriminative features within each frame and model contextual relationships among frames. The dynamic-static fusion module [21, 22] is used to obtain more robust and discriminative spatial features from both static features and dynamic features, which can effectively reduce the interference of noisy frames on the DFER task. In addition, Wang *et al*. [47] propose the Dual Path multi-excitation Collaborative Network (DPCNet) to learn critical information for facial expression representation from fewer key frames.

The methods mentioned above approach DFER as a general video understanding task and do not consider the weakly supervised nature of the problem due to inexact crowd-sourced annotation. Moreover, they overlook the issue of imbalanced short- and long-term temporal relationships in DFER and rely solely on a sequence model. By contrast, the proposed M3DFEL framework addresses these challenges at their root by tackling the weakly supervised problem and modeling the imbalanced short- and long-term temporal relationships in a unified manner.

2.2. Multi-Instance Learning

MIL is a technique designed to address the inexact labeling problem [8]. Traditionally, each sample is treated as a bag of instances, where the bag is labeled negative only when all instances are negative. Otherwise, the bag is considered positive. MIL is commonly used in scenarios where there are a large number of samples with only one label. In these situations, the methods must accurately identify and recognize positive instances within a dataset that contains a significant proportion of negative instances.

MIL has been applied in various fields, such as WSOD (weakly supervised object detection) [9, 39], action location [30], and WSI (whole slide image) classification [49, 53]. Although there is no research that formulates in-the-wild DFER as a MIL problem, we can draw insights from WSOD methods, which also solve the MIL problem in video-based tasks. For instance, Feng *et al*. [7] propose an end-to-end weakly supervised Rotation-Invariant Aerial Object Detection Network to tackle object rotations without corresponding constraints. Meanwhile, Tang *et al*. [39] introduce a novel online instance classifier refinement algorithm to integrate MIL and the instance classifier refinement procedure into a single deep network, and train the network end-to-end with only image-level supervision.

The use of MIL has been explored in recognizing emotions. Romeo *et al*. [33] explores the usage of some existing MIL-based SVMs in detecting the emotion using physiological signals. Chen *et al*. [2] mainly focuses on Action Unit encoding for pain detection, and applies clustering-based maximum operation for instance fusion in MIL. Wu *et al*. [48] employ a differentiable OR operation for MIL with Hidden Markov Model as the classifier in lab-controlled DFER, using facial landmark as input feature. All of these methods use handcrafted features and employ conventional machine learning MIL methods on their tasks. Moreover, the samples of lab-controlled DFER are more unambiguous and the environment and facial expression dynamics are fixed, while the in-the-wild samples are more complex and challenging. Their ways of applying the MIL method are not applicable to our situation with in-the-wild DFER.

With high-level hypotheses and observations to DFER, we design our novel MIL framework through fusing the modeling of the imbalanced temporal relationship within the MIL pipeline. In contrast to using existing MIL methods to fuse the handcrafted feature, we model the strong short-term temporal relationship during feature extraction and learn the long-term relationship during instance fusion.

3. Method

3.1. Overview

The MIL pipeline typically involves four steps: Instance Generation, Instance Feature Extraction, Instance Aggregation, and Classification. In the case of DFER, the proposed M3DFEL framework follows this pipeline and utilizes 3DCNNs to extract features from the generated 3D-instances and learn the short-term temporal relationship. The DLIAM is used to model the long-term temporal relationship while dynamically fusing the instances into a bag. To maintain temporal consistency on both the bag-level and instance-level, the DMIN is introduced. An overview of the proposed M3DFEL framework is illustrated in Figure 2.

3.2. Proposed Method

Three-Dimensional Instance Generation. Generating instances by cropping a video into frames is a common approach for MIL tasks, as they are usually frame-based tasks such as weakly supervised object detection or action location. However, in DFER, some frames may not capture a typical facial expression when the subject is talking. While such frames appear abnormal on their own, they actually represent the motion of facial movement. Additionally, compared to other MIL tasks, the differences in facial movements between classes are subtle, which means that even small movements can cause changes in the predicted emotion and features.

The proposed 3D-Instance Generation addresses these problems with a simple yet effective approach. Given a video V that contains T frames of images, we crop a video into N parts in dimension T. Then, the bag can be defined as a sequence of instances $\mathrm{I} = [I_1, I_2, ..., I_N]$, where $I_n \in \mathbb{R}^{C \times T \times H \times W}$ denotes the n-th 3D instance. This design enables the feature extractor to model the strong short-term temporal relationship by capturing the motion of facial movement across the instances, as well as the consistent emotion when the subject is talking. This is crucial in DFER, where the facial movements and emotional differences are subtle, and even tiny movements can significantly

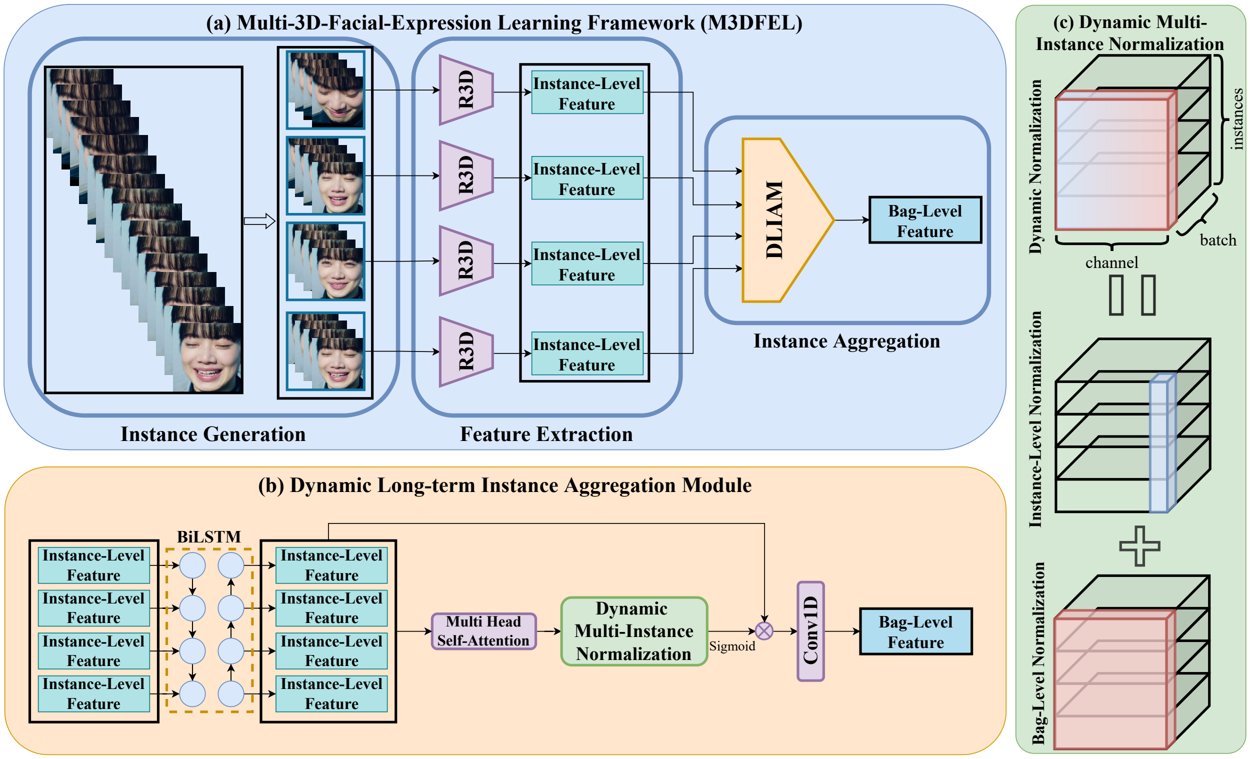


Figure 2. An overview of the proposed M3DFEL framework. (a) The pipeline of M3DFEL: Three-Dimensional Instance Generation, Instance Feature Extraction, Long-term Instance Aggregation and Classification. (b) The structure of the proposed DLIAM. (c) The sketch of Dynamic Multi-Instance Normalization (DMIN).

impact the predicted emotion.

Instance Feature Extraction. The vanilla R3D18 is used to extract the feature F_n for each instance I_n within the bag. The R3D18 model extracts compressed frame representations and incorporates the temporal information of neighboring frames for each instance. This results in a bag of feature representations for the instances, denoted as $\mathrm{F} \in \mathbb{R}^{N \times C}$, where C represents the number of channels.

Dynamic Long-term Instance Aggregation. As aforementioned, there exists an imbalance between long- and short-term temporal relationships in DFER. As the 3D-Instance-based MIL setting strengthens the short-term temporal learning, the Dynamic Long-term Instance Aggregation Module (DLIAM) is proposed to dynamically aggregate the instances while modeling the long-term temporal relationship. The first step is to use a BiLSTM to capture the long-term temporal relationship between instances.

After that, to dynamically aggregate the representations of instances, we first apply the Multi-Head Self-Attention (MHSA) to learn the inter-instance relationship and obtain an attention weight $\mathrm{A} \in \mathbb{R}^{N \times C}$.

Moreover, we find out that the recognition results of the instances are rather unstable, which violates the common sense that the emotion status is relatively stable and continuous in a short period of time, such as a few seconds. To address this concern, we draw inspiration from [29] and design a Dynamic Multi-Instance Normalization (DMIN) method to maintain temporal consistency at both the bag and instance levels. We define a set of normalizers $\mathcal{K} = bn, in$ and dynamically adjust the importance weights, where bn denotes the bag-level normalizer and in denotes the instance-level normalizer. Let A_{nc} and $\hat{A}_{nc}$ be the c-th channel value of the n-th instance before and after normalization, and the normalization procedure can be presented as follows,

$$\hat{A}_{nc} = \frac{A_{nc} - \sum_{k \in \mathcal{K}} w_k \mu_k}{\sqrt{\sum_{k \in \mathcal{K}} w'_k \sigma_k^2 + \epsilon}} * \gamma + \beta, \tag{1}$$

where μ_k and σ_k are the mean and variance values, respectively, estimated using the normalizer k for the specific channel value of the instance. ϵ is a small number added for numerical stability. The learnable affine transform parameters are represented by γ and β. The importance weights of the normalizer k are represented by w_k and w'_k, and are dynamically adjusted.

The difference between the two normalizers is the value set to estimate the statistics. The bag-level normalization computes the statics along the dimension of N and C for each bag,

$$\mu_{bn} = \frac{1}{NC}\sum_{n,c}^{N,C} A_{nc}, \quad \sigma_{bn} = \frac{1}{NC}\sum_{n,c}^{N,C}(A_{nc} - \mu_{bn})^2, \tag{2}$$

where $\mu_{bn}, \sigma_{bn} \in \mathbb{R}^1$, suggesting that the values of the single bag share the same bag-level statistics. The instance-level normalization computes the statics across the dimension of N,

$$\mu_{in} = \frac{1}{N}\sum_{n}^{N} A_{nc}, \quad \sigma_{in} = \frac{1}{N}\sum_{n}^{N}(A_{nc} - \mu_{in})^2, \tag{3}$$

where $\mu_{in}, \sigma_{in} \in \mathbb{R}^C$, suggesting that the instance-level statistics are shared within the same channel of each bag.

For the importance weights w_k and w'_k, we use the softmax operation to ensure $\sum_{k\in\mathcal{K}} w_k = 1$, $\sum_{k\in\mathcal{K}} w'_k = 1$ and the scalars are restricted between 0 and 1,

$$w_k = \frac{e^{\lambda_k}}{\sum_{j\in\mathcal{K}} e^{\lambda_j}}, \tag{4}$$

where λ is the learnable parameter to adjust the weights for different normalization approaches.

For the final aggregation of the instances, the weights are first multiplied with the instances after a sigmoid function. Then, a Conv1D layer is utilized to aggregate the instance-level features X into bag-level feature $\mathrm{Z} \in \mathbb{R}^{N\times C}$,

$$\mathrm{Z} = Conv1D(\mathrm{X} * Sigmoid(\hat{\mathrm{A}})). \tag{5}$$

The bag-level feature is then fed into a fully connected layer to obtain the prediction result, and a Cross Entropy Loss is used to supervise the results.

4. Experiments

4.1. Datasets

DFEW [11] is a large-scale in-the-wild dataset introduced in 2020, which contains over 16,000 video clips with dynamic facial expressions. These clips are collected from more than 1,500 movies worldwide, and they contain various challenging interferences, such as extreme illuminations, self-occlusions, and unpredictable pose changes. Each video clip is annotated individually by ten well-trained annotators under professional guidance and assigned to one of the seven basic expressions, including *Happy*, *Sad*, *Neutral*, *Angry*, *Surprise*, *Disgust*, and *Fear*. We adopt the 5-fold cross-validation setting provided by DFEW to ensure a fair comparison among different methods.

FERV39K [46] is currently the largest in-the-wild DFER dataset, comprising 38,935 video clips collected from 4 scenarios, which are further subdivided into 22 fine-grained scenes. It is the first DFER dataset with a large-scale number of 39K clips, scenario-scene division, and cross-domain supportability. Each video clip in FERV39K is annotated by 30 professional annotators to ensure high-quality labels and assigned to one of the seven primary expressions as in DFEW. We use the training and testing sets provided by FERV39K for fair comparison.

4.2. Implementation Details

Our entire framework is implemented using PyTorch-GPU and trained on Tesla V100 GPUs. For feature extraction, we employ the vanilla R3D18 model and utilize its pre-trained weights provided by Torchvision. The models are trained for 300 epochs with 20 warm-up epochs using the AdamW optimizer and cosine scheduler. The learning rate is set to 5e-4, the minimum learning rate is set to 5e-6, and the weight decay is set to 0.05. We use a batch size of 256 and apply label smoothing with a value of 0.1. Our augmentation techniques consist of random cropping, horizontal flipping, and 0.4 color jitter. For each video, we extract a total of 16 frames as our sample. In all experiments, we use weighted average recall (WAR) and unweighted average recall (UAR) as evaluation metrics, with more emphasis placed on the WAR as it is considered to be the critical metric. In the following experiments, we focus on using DFEW [11] for further analysis and discussion.

4.3. Comparison with the State-of-the-art Methods

We compare our method with the state-of-the-art methods on two in-the-wild datasets DFEW and FERV39K.

DFEW. The results, obtained under 5-fold cross-validation, are presented in Table 1. It can be observed that the proposed M3DFEL achieves the best performance in terms of both WAR and UAR, using vanilla R3D18 as the backbone. The results are better than those obtained with NR-DFERNet [22], with a difference of 1.06% in terms of WAR and 1.89% in terms of UAR. The performance of M3DFEL on each expression is also shown in Table 1, and more detailed analysis is presented in Section 4.5.

FERV39K. The results are shown in Table 2. FERV39K is a challenging DFER dataset, resulting in lower overall accuracy compared to DFEW. M3DFEL outperforms NR-DFERNet [22] by 1.70%/1.95% of WAR/UAR. Notably, using vanilla R3D18 and LSTM, M3DFEL significantly surpasses 3DResNet18 [10] and R18+LSTM [46] by 10.10%/9.27% and 4.72%/5.02% of WAR/UAR, which indicates the effectiveness of the framework.

4.4. Ablation Study

Evaluation of different bag sizes. We conduct ablation studies on DFEW to demonstrate the impact of bag size in the MIL setting. When the bag size is set to 16, the same as the number of sampled frames, the 3DMIL setting degrades

Method	Accuracy of Each Emotion(%)							Metrics(%)		FLOPs(G)
	Hap.	Sad	Neu.	Ang.	Sur.	Dis.	Fea.	WAR	UAR	
C3D [40]	75.17	39.49	55.11	62.49	45.00	1.38	20.51	53.54	42.74	38.57
P3D [32]	74.85	43.40	54.18	60.42	50.99	0.69	23.28	54.47	43.97	-
I3D [1]	78.61	44.19	56.69	55.87	45.88	2.07	20.51	54.27	43.40	6.99
R(2+1)D18 [41]	79.67	39.07	57.66	50.39	48.26	3.45	21.06	53.22	42.79	42.36
3D ResNet18 [10]	73.13	48.26	50.51	64.75	50.10	0.00	26.39	54.98	44.73	8.32
ResNet18+LSTM [11]	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	53.08	42.86	7.78
EC-STFL [11]	79.18	49.05	57.85	60.98	46.15	2.76	21.51	56.51	45.35	8.32
FormerDFER [54]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	65.70	53.69	9.11
STT [31]	87.36	67.90	64.97	71.24	53.10	3.49	**34.04**	66.45	54.58	-
DPCNet [47]	-	-	-	-	-	-	-	66.32	55.02*	9.52
NR-DFERNet [22]	88.47	64.84	**70.03**	**75.09**	**61.60**	0.00	19.43	68.19	54.21	6.33
M3DFEL(Ours)	**89.59**	**68.38**	67.88	74.24	59.69	0.00	31.63	**69.25**	**56.10**	**1.65**

Table 1. Comparison(%) of our M3DFEL with the state-of-the-art methods on DFEW. * indicates the result is calculated according to the confusion matrix reported in the paper. (Bold: Best result, Underline: Second best)

Method	WAR	UAR
C3D [40]	31.69%	22.68%
P3D [32]	33.39%	23.20%
I3D [1]	38.78%	30.17%
R(2+1)D18 [41]	41.28%	31.55%
3D ResNet18 [10]	37.57%	26.67%
R18+LSTM [46]	42.95%	30.92%
2R18+LSTM [46]	43.20%	31.28%
NRDFERNet [22]	45.97%	33.99%
M3DFEL(Ours)	**47.67%**	**35.94%**

Table 2. Comparison(%) of our M3DFEL with the state-of-the-art methods on FERV39K.

Bag size	WAR	UAR
1	68.04%	55.36%
2	68.55%	55.92%
4	**69.25%**	**56.10%**
8	68.24%	55.32%
16	66.36%	53.56%

Table 3. The ablation Study of different bag sizes. The video sample rate is 16. Bag size 1 indicates that the entire sampled video is fed into the feature extractor, rendering the MIL pipeline and instance aggregation module inapplicable. Bag size 16 denotes that each instance consists of a single frame.

Setting	Module	WAR	UAR
a	baseline	68.23	55.44
b	w/o LSTM	68.63	55.62
c	w/o DMIN	68.91	56.03
d	w/o MHSA	69.13	**56.21**
e	M3DFEL	**69.25**	56.10

Table 4. Ablation Study of the proposed Dynamic Long-term Instance Aggregation Module. DMIN is the abbreviation of Dynamic Multi-Instance Normalization. MHSA is the abbreviation of Multi-Head Self-Attention.

to 2D, where ResNet18 is used as the backbone. Setting the bag size to 1 represents feeding all frames into the feature extractor at once, leading to a normal supervised learning paradigm where the aggregation module fails. The results are shown in Table 3. When the 3DMIL setting degrades to 2D with the bag size of 16, the model achieves a WAR of 66.36% and a UAR of 53.56%. This setting has a large margin compared to the others, probably because DLIAM only learns a weak temporal relationship and lacks modeling the strong temporal relationship. Although it is essential to learn a strong temporal relationship, the experiment with a bag size of 1 shows that it is not always the best solution. With a WAR of 68.04% and a UAR of 55.36%, using the whole video as input falls behind the other 3DMIL settings. The results show that, with a sample rate of 16, setting the bag size to 4, where each instance contains four frames, is an appreciable choice.

We conduct additional experiments to analyze the classification performance on a single instance. As depicted in Figure 3, when the subject expressed emotions through subtle facial movements, the 3D-instance-based MIL model was able to capture these movements and make accurate predictions. In contrast, the 2D-instance-based MIL model only succeeded in a few frames.

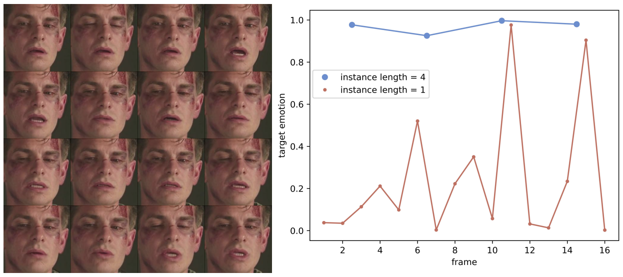
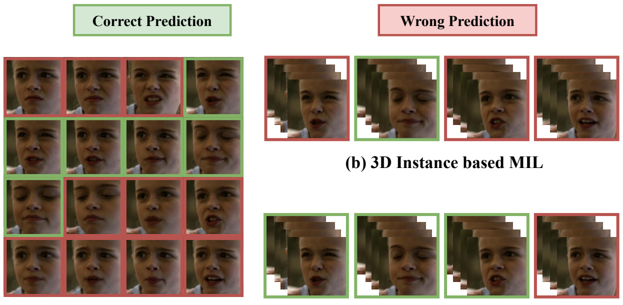


Figure 3. Evaluation of 2D-instance-based MIL and 3D-instance-based MIL. The target emotion is *Sad*.



(b) 3D Instance based MIL

(a) 2D Instance based MIL

(c) M3DFEL

Figure 4. Visualization results of different MIL methods.

Evaluation of the DLIAM. We conduct an experiment to study the effectiveness of the proposed DLIAM on DFEW, using Average Pooling as the baseline method. The results are presented in Table 4. The comparison shows that using an attention-based method to aggregate the instances, as done in existing MIL methods, performs worse than the BiLSTM-based settings. Furthermore, the results demonstrate that the proposed plug-and-play DMIN improves performance at a fraction of the cost. The complete DLIAM setting (e) outperforms the baseline by 1.02%/0.66% of WAR/UAR on DFEW, fully indicating the effectiveness of DLIAM and the importance of long-term temporal relationship modeling in DFER.

4.5. Visualization

To further evaluate the effectiveness of our method, visualization studies are conducted.

Visualization of different MIL methods. To investigate how M3DFEL works, we obtain the classification result on a single instance in a sample with strong facial movements. As shown in Figure 4, the 2D-instance-based MIL is greatly influenced by the facial movements when the subject is talking. At the same time, a simple 3D-instance-based MIL captures the information that the subject is talking happily in the second instance, but it still predicts the other instances as other non-target emotions. With the DLIAM, M3DFEL can further recognize the confusing expressions with the confidence given by the second instance, and then successfully predicts the emotional status of these confusing samples.

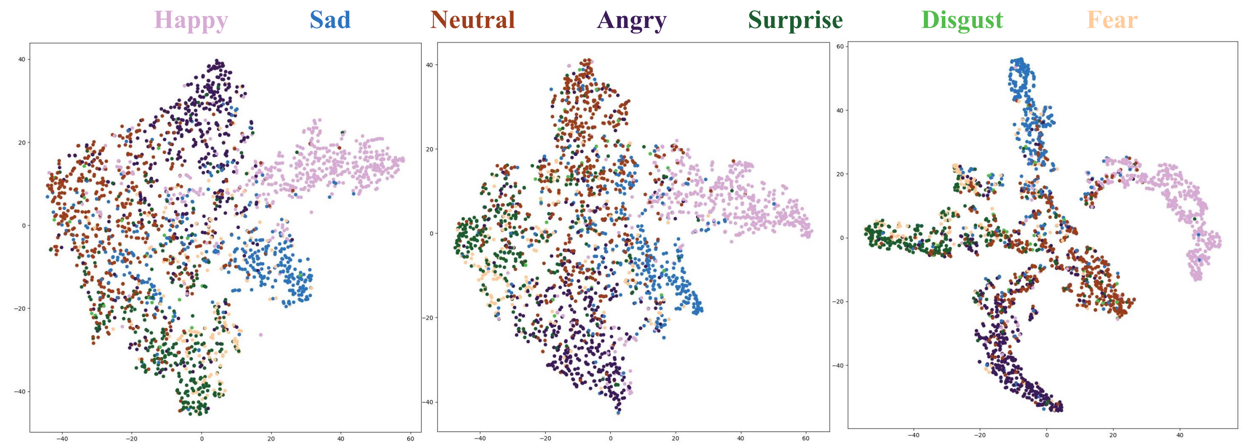
T-SNE Visualization. We utilize t-SNE [42] to visualize the distribution of dynamic facial expression features extracted by our baselines and M3DFEL. The t-SNE plot in Figure 5 illustrates that the features obtained by the baselines lack discriminative power, with a significant overlap between categories. In contrast, our proposed M3DFEL method shows a clearer boundary between categories, with more concentrated clusters. Nonetheless, the t-SNE figure indicates that many instances of the *Neutral* expression are present in other emotions, and some other emotions may also be classified as *Neutral*. In DFER, many expressions have lower intensity than in SFER, and annotators may have access to more information and be more confident with these micro-expressions. However, recognizing these low-intensity expressions is a challenging task for the model, leading to difficulty in distinguishing *Neutral* and low-intensity expressions.

Confusion matrix. We visualize the confusion matrix of our proposed M3DFEL evaluated on DFEW Fold 1 5 to analyze the results. From Figure 6, we observe that the model struggles to predict the emotion of videos labeled as *Disgust*. This is due to the severe label imbalance in DFEW, where the proportion of *Disgust* videos is only 1.22%. As a result, the model is more likely to ignore videos with label *Disgust* during training, leading to poor performance on this emotion. A similar situation occurs for the *Fear* label, which has a proportion of 8.14%. The model tends to predict some videos with label *Fear* as other emotions, due to the lack of sufficient training examples for this emotion. Additionally, we observe that the model tends to predict the label *Neutral* more frequently. This is because predicting these samples as any other emotion is more risky than predicting them as *Neutral*.

5. Discussion

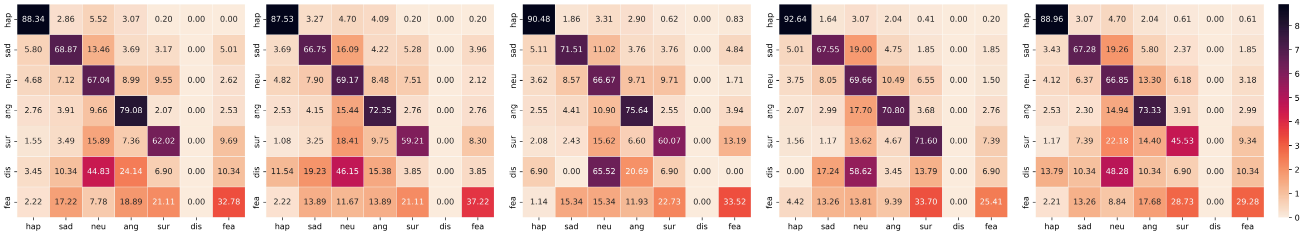
It is clear that there are still many unresolved issues in DFER, despite the proposed M3DFEL framework. Our analysis of the failure cases reveals that most of them occur during the classification stage rather than instance fusion in MIL. For example, if the majority of the video frames are neutral, the fusion result of the whole bag is the non-neutral emotion, as expected. However, the model often misclassifies the non-neutral emotion, e.g. classifying a *Fear* as a *Surprise*. This indicates that the current performance is largely limited by the model's classification ability.

One major issue is the imbalanced label problem, where the accuracy on *Disgust* and *Fear* is sacrificed due to the lack of samples with these labels in the dataset. This problem is more severe in DFER than in SFER, indicating that solely utilizing DFER datasets may be insufficient. Possi-



(a) 2D-instance-based MIL (b) 3D-instance-based MIL (c) M3DFEL

Figure 5. 2D t-SNE visualization [42] of dynamic facial expression features obtained by different MIL methods, including 2D-instance-based MIL, 3D-instance-based MIL and M3DFEL. The features are extracted from the DFEW dataset.



(a) Fold 1 (b) Fold 2 (c) Fold 3 (d) Fold 4 (e) Fold 5

Figure 6. The confusion matrix of our proposed M3DFEL evaluated on DFEW Fold 1-5.

ble solutions to this issue include transfer learning or self-supervised pre-training methods. Another issue is that some expressions in DFER have much lower intensity than static expressions, which is similar to the key problem in micro-expression recognition (MER). Utilizing MER techniques such as optical flow may help to solve this problem. Additionally, some prior knowledge like landmarks or Action Units may provide useful hints to the model. Except for these issues, the noisy label problem, the uncertainty problem and the hard sample problem all influence DFER greatly. More importantly, it is difficult for us to distinguish if we should emphasize or weaken the learning of the samples. Beyond the existing problems, we hope that the model should not overfit on the dataset itself. As FERV39K provides cross-domain supportability, domain-generalization is an important research direction.

6. Conclusion

In this study, we conduct a thorough analysis of the DFER problem and proposed a new learning paradigm. We utilize the Multi-Instance Learning (MIL) pipeline and develop the M3DFEL framework to address the weakly supervised problem and imbalanced short- and long-term temporal relationships in a unified manner. The M3DFEL framework includes the 3D-Instance Generation module, which learns the strong short-term temporal relationship, and the Dynamic Long-term Instance Aggregation Module (DLIAM), which models weak long-term temporal relationships. The proposed framework also implements dynamic normalization to maintain temporal consistency at both bag-level and instance-level. Our extensive experiments support our perspective on the DFER problem and demonstrated the effectiveness of the proposed M3DFEL framework. In addition, we have identified several research directions that may guide future studies in this field, such as the imbalanced label problem, the uncertainty problem, and others.

7. Acknowledgments

This research is funded by the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901).

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[2] Zhanli Chen, Rashid Ansari, and Diana J. Wilkie. Learning pain from action unit combinations: A weakly supervised approach via multiple instance learning. *IEEE Transactions on Affective Computing*, 13(1):135–146, 2022. 3

[3] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: Deformable patch attack. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022*. Springer, 2022. 2

[4] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 15148–15158, June 2022. 2

[5] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016. 2

[6] Zhixin Fang, Libai Cai, and Gang Wang. Metahuman creator the starting point of the metaverse. In *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*, pages 154–157. IEEE, 2021. 1

[7] Xiaoxu Feng, Xiwen Yao, Gong Cheng, and Junwei Han. Weakly supervised rotation-invariant aerial object detection network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14146–14155, 2022. 3

[8] Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential. *arXiv preprint arXiv:2206.04425*, 2022. 3

[9] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervison. *arXiv preprint arXiv:2203.11652*, 2022. 3

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 5, 6

[11] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng 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. 1, 5, 6

[12] Ziv Lautman and Shahar Lev-Ari. The use of smart devices for mental health diagnosis and care, 2022. 1

[13] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019. 2

[14] Min Kyu Lee, Dong Yoon Choi, Dae Ha Kim, and Byung Cheol Song. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8, 2019. 2

[15] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *The Thirty-Third AAAI Conference*, 2019. 2

[16] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *2019 IEEE/CVF International Conference on ICCV*, pages 8518–8527. IEEE, 2019. 2

[17] Bo Li, Zhengxing Sun, Lv Tang, and Anqi Hu. Two-b-real net: Two-branch network for real-time salient object detection. In *IEEE International Conference on ICASSP*. IEEE, 2019. 2

[18] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi. Detecting robust co-saliency with recurrent co-attention neural network. In Sarit Kraus, editor, *IJCAI*, 2019. 2

[19] Bo Li, Zhengxing Sun, Quan Wang, and Qian Li. Co-saliency detection based on hierarchical consistency. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on MM*, pages 1392–1400. ACM, 2019. 2

[20] Bo Li, Lv Tang, Senyun Kuang, Mofei Song, and Shouhong Ding. Toward stable co-saliency detection and object co-segmentation. *IEEE Trans. Image Process.*, 31:6532–6547, 2022. 2

[21] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2208.10335*, 2022. 1, 2, 3

[22] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022. 1, 2, 3, 5, 6

[23] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4166–4175, 2022. 1

[24] Zuojin Li, Liukui Chen, Ling Nie, and Simon X Yang. A novel learning model of driver fatigue features representation for steering wheel angle. *IEEE Transactions on Vehicular Technology*, 71(1):269–281, 2021. 1

[25] Feng Liu, Si-Yuan Shen, Zi-Wang Fu, Han-Yang Wang, Ai-Min Zhou, and Jia-Yin Qi. Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition. *Entropy*, 24(7):1010, 2022. 1

[26] Feng Liu, Hanyang Wang, Jiahao Zhang, Ziwang Fu, Aimin Zhou, Jiayin Qi, and Zhibin Li. Evogan: An evolutionary computation assisted gan. *Neurocomputing*, 469:81–90, 2022. 1

[27] Feng Liu, Han-Yang Wang, Si-Yuan Shen, Xun Jia, Jing-Yi Hu, Jia-Hao Zhang, Xi-Yi Wang, Ying Lei, Ai-Min Zhou, Jia-Yin Qi, et al. Opo-fcm: A computational affec-

tion based occ-pad-ocean federation cognitive modeling approach. *IEEE Transactions on Computational Social Systems*, 2022. 1

[28] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 646–652, New York, NY, USA, 2018. Association for Computing Machinery. 2

[29] Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. In *International Conference on Learning Representations*, 2018. 4

[30] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *European conference on computer vision*, pages 729–745. Springer, 2020. 3

[31] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022. 1, 2, 6

[32] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 6

[33] Luca Romeo, Andrea Cavallo, Lucia Pepa, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multiple instance learning for emotion recognition using physiological signals. *IEEE Transactions on Affective Computing*, 13(1):389–407, 2022. 3

[34] Siyuan Shen, Feng Liu, and Aimin Zhou. Mingling or misalignment? temporal shift for speech emotion recognition with pre-trained representations. *arXiv preprint arXiv:2302.13277*, 2023. 1

[35] Lv Tang and Bo Li. CLASS: cross-level attention and supervision for salient objects detection. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *ACCV 2020*, 2020. 2

[36] Lv Tang and Bo Li. Cosformer: Detecting co-salient object with transformers. *arXiv preprint arXiv:2104.14729*, 2021. 2

[37] Lv Tang, Bo Li, Senyun Kuang, Mofei Song, and Shouhong Ding. Re-thinking the relations in co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2

[38] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *2021 IEEE/CVF ICCV*, pages 3560–3570. IEEE, 2021. 2

[39] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2843–2851, 2017. 3

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6

[41] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7, 8

[43] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 1

[44] Weijie Wang, Nicu Sebe, and Bruno Lepri. Rethinking the learning paradigm for facial expression recognition. *arXiv preprint arXiv:2209.15402*, 2022. 1

[45] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022. 1

[46] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang 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. 1, 5, 6

[47] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 101–110, 2022. 1, 2, 3, 6

[48] Chongliang Wu, Shangfei Wang, and Qiang Ji. Multi-instance hidden markov model for facial expression recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6, 2015. 3

[49] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 3

[50] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. DENSE: Data-free one-shot federated learning. In *Advances in NeurIPS*, 2022. 2

[51] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *Proceedings of the ICML*. PMLR, 2022. 2

[52] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 49(3):839–847, 2018. 2

[53] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020. 3

[54] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 1, 2, 6

[55] Jiawen Zheng, Bo Li, ShengChuan Zhang, Shuang Wu, Liujuan Cao, and Shouhong Ding. Attack can benefit: An adversarial approach to recognizing facial expressions under noisy annotations. In *AAAI Conference*, 2023. 1

[56] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 4504–4513, June 2022. 2

[57] Yijie Zhong, Bo Li, Lv Tang, Hao Tang, and Shouhong Ding. Highly efficient natural image matting. *CoRR*, abs/2110.12748, 2021. 2