Supplementary Materials for "MAFW"

A ADDITIONAL DETAILS IN MAFW

A.1 Multiple Expression Distribution

Table 1 shows the distributions of 32-class multiple emotion categories on the multiple expression set, including the clip length and clip amount.

Table 1: The distributions of clip amount and clip length per multiple emotion category on the multiple expression subset.

Expressions	Clips				Percent(%)
	0-2s	2-5s	5s+	Total	
Anger,Disgust	125	601	157	883	21.76
Sadness,Helplessness	21	307	214	542	13.36
Fear,Surprise	112	252	28	392	9.66
Surprise,Anxiety	36	173	23	232	5.72
Fear,Anxiety	37	146	35	218	5.37
Disgust,Contempt	26	151	31	208	5.13
Happiness,Surprise	23	152	31	206	5.08
Sadness,Anxiety	21	113	51	185	4.56
Anxiety,Helplessness	14	104	38	156	3.84
Disgust,Anxiety	16	96	19	131	3.23
Helplessness,Disappointment	10	64	24	98	2.41
Fear,Sadness	13	69	14	96	2.37
Anger,Sadness	9	55	26	90	2.22
Sadness,Anxiety,Helplessness	3	47	26	76	1.87
Anxiety,Helplessness,Disappointment	4	39	16	59	1.45
Anger,Anxiety	7	38	9	54	1.33
Happiness,Contempt	3	42	7	52	1.28
Fear,Surprise,Anxiety	8	27	10	45	1.11
Disgust,Helplessness	5	24	10	39	0.96
Disgust,Surprise	13	20	4	37	0.91
Anger,Disgust,Anxiety	3	31	2	36	0.89
Anger,Surprise	6	25	2	33	0.81
Sadness,Disappointment	2	21	9	32	0.79
Sadness,Surprise	1	14	11	26	0.64
Sadness,Helplessness,Disappointment	0	11	11	22	0.54
Fear,Sadness,Anxiety	4	13	2	19	0.47
Disgust,Anxiety,Helplessness	4	13	2	19	0.47
Anger,Disgust,Contempt	2	11	5	18	0.44
Disgust,Sadness	2	12	3	17	0.42
Anger,Helplessness	3	7	3	13	0.32
Disgust,Disappointment	3	7	2	12	0.30
Disgust,Helplessness,Disappointment	3	9	0	12	0.30
Total	539	2694	825	4058	100.00

A.2 Annotation Format of the Compound Emotion

To efficiently annotate compound emotions, we developed an annotation tool called ExprLabelTool to generate and save annotation files for each annotator. Fig. 1 shows an annotated file format of a video-audio clip in MAFW. The "video_id" represents the index of the video-audio clip, the "labels" represents the expression categories labeled by an annotator for the clip, and the "scores" represents the self-confidence scores corresponding to the expression categories.

```
{
    "video_id" : 05237.mp4,
    "labels"   : [Disgust, Contempt],
    "scores"   : [0.6, 1.0]
}
```

Figure 1: An example of the emotion annotation file in MAFW.

A.3 Annotation Format of the Emotional Descriptive Text

We carefully design our caption annotation task for emotional descriptive texts and develop several rules to ensure the sentences are of high syntactic and semantic quality in MAFW. Table. 2 shows the annotation instructions given to the annotators for the emotional description text.

Table 2: The annotation instructions given to the annotator for the emotional description text.

Task
The task is to describe the emotional elements and the movements of the five facial features of the only main character in the video. The emotional elements include the body actions, the environment, the persons the character is speaking to, the tone of voice, and the the events' context.
DOs
1. Each emotional description text is available in both Chinese and English. 2. Use a personal pronoun as the subject of the sentence to refer to the main character in the video, such as "an old man", "a boy", etc., rather than their names (either the character's name or the actor's name). 3. Use the simple present tense. 4. Try to describe the part of the emotional elements in one sentence and modify the verb with an appropriate adverb to emphasize the sentiment state of the character. This part should be no less than eight words. 5. Use predefined sentences to describe the part of the five facial features without adding new descriptions arbitrarily. 6. Each sentence should be grammatically correct. 7. It should describe only the visual content in the video.
DONTs
1. Words that directly specify the expression category, such as "angry/anger/angrily", "sad/sadness/sadly", etc., should **NOT** appear. 2. It should **NOT** describe your opinions, guesses or subjective judgements. 3. It should **NOT** contain any digits.

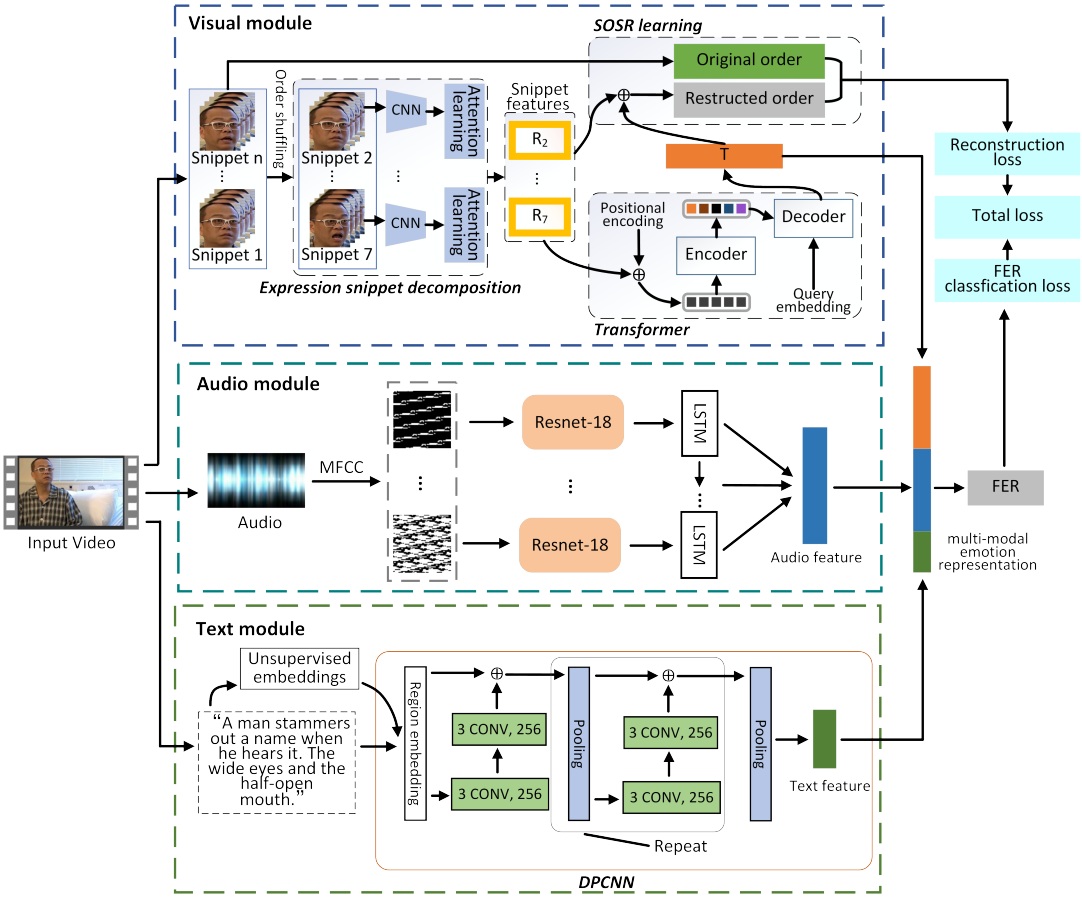

Figure 2: The framework with the T-ESFL for multi-modal emotion recognition.

B THE DETAILED FRAMEWORK FOR MULTI-MODAL EMOTION RECOGNITION

The detailed framework with the T-ESFL for multi-modal emotion recognition is shown in Fig. 2. The multi-modal T-ESFL consists of three main modules, namely the visual module, the audio module, and the text module. First, the visual module uses snippet-based Transformer and SSOR to obtain the salient emotion feature T, the audio module uses ResNet_LSTM [2–4] to extract the audio emotion feature, and the text module uses DPCNN [5] to extract the text emotion feature. Then, we concatenate the visual, audio, and text features to generate the multi-modal emotion representation. As with the single-modal T-ESFL, the total objective function in the multi-modal emotion recognition includes cross-entropy loss and the snippet order reconstruction loss.

C EXPERIMENTS FOR VIDEO EMOTIONAL CAPTIONING

We further discuss the application of our database to another task, *e.g.*, video emotional captioning. To this end, we used two off-the-shelf video captioning models, namely Reconstruction network[10] and Video paragraph captioning model [8], to perform the video emotional captioning task and generate emotional text descriptions.

Table 3: Performance evaluation of video emotional captioning on our MAFW.

Model	BLEU-4	METEOR	CIDEr
Reconstruction network [10]	6.87	11.50	25.63
Video paragraph captioning model [8]	9.09	15.49	23.40

We used three widely-used standard metrics in video captioning to evaluate the generated emotional text descriptions, namely BLEU-4[7], METEOR[1], and CIDEr[9]. Table 3 shows the experimental results of video emotional captioning using these two models in our database. Additionally, qualitative examples for video emotional captioning are shown in Figure 3.

D ETHICAL STATEMENT

Although this is a purely academic investigation, the potential sensitivity of facial information necessitates an explicit statement of the ethics involved.

Privacy. Our method is used to capture features of facial expressions shared by many individuals, which are related to the common human perception of expressions. Therefore our method

Reconstruction network: A man speaks to the camera, a man talks loudly
Video paragraph captioning model: A man speaks to the man in front of him The slight frown
Ground Truth: A man talks loudly. The tight frown and a downward pull on the right lip corner.

Reconstruction network: A woman speaks to the man in front of her, a woman talks to a man and expresses her displeasure
Video paragraph captioning model: A woman is not satisfied with the man in front of her The wide eyes the wrinkled nose
Ground Truth: A woman talks to a man and expresses her displeasure. The wide eyes, the higher inner corners of eyebrows and the lower outer corners of eyebrows.

Figure 3: Visualization examples of video emotional captioning. The words in red are the predicted results of each model close to the Ground Truth, and the words in green are the Ground Truth.

does not produce individual-specific facial expression analysis. Our MAFW database is used for academic research only and is compliant with GDPR [1] principles. The copyright of the original and cropped versions of the video remains with the original owner. No commercialization, secondary distribution or alteration of MAFW is allowed by any applicant.

Database Bias. During the data collection process, we did not differentiate any factors like gender, race, geography, age, etc. However, some data bias may occur in our MAFW database due to objective limitations such as data sources, the difficulty of collecting different emotions, etc.

Metadata. In our MAFW metadata, we use only the gender statistics automatically inferred from the model pre-trained on CelebA[6]. We only use this information to evaluate the distribution of data in our MAFW database and do not make use of it in our experiments or elsewhere.

REFERENCES

[1] Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*. The Association for Computer Linguistics, 376–380. https://doi.org/10.3115/v1/w14-3348

[2] F.A. Gers, J. Schmidhuber, and F. Cummins. 1999. Learning to forget: continual prediction with LSTM. In *9th International Conference on Artificial Neural Networks (ICANN)*, Vol. 2. 850–855. https://doi.org/10.1049/cp:19991218

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[5] Rie Johnson and Tong Zhang. 2017. Deep Pyramid Convolutional Neural Networks for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. ACL, 562–570. https://doi.org/10.18653/v1/P17-1052

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*. 3730–3738. https://doi.org/10.1109/ICCV.2015.425

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[8] Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards Diverse Paragraph Captioning for Untrimmed Videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 11245–11254. https://openaccess.thecvf.com/content/CVPR2021/html/Song_Towards_Diverse_Paragraph_Captioning_for_Untrimmed_Videos_CVPR_2021_paper.html

[9] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

[10] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018. Reconstruction Network for Video Captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 7622–7631. https://doi.org/10.1109/CVPR.2018.00795

[1] https://gdpr-info.eu/