# Sequential Interactive Biased Network for Context-Aware Emotion Recognition

Xinpeng Li[1,2] Xiaojiang Peng[1] * Changxing Ding[2] *
[1] Shenzhen Technology University [2] South China University of Technology
li.xin.peng@outlook.com pengxiaojiang@sztu.edu.cn chxding@scut.edu.cn

## Abstract

*Emotion context information is crucial yet complicated for emotion recognition. How to process it is a challenging problem. Existing works mainly extract context representations of the face, body and scene independently. These strategies may be limited in the understanding of emotional context relation. To address this problem, we propose Sequential Interactive Biased Network (SIB-Net), which is motivated by the studies that the context contains sequential, interactive and biased relation. Specifically, SIB-Net captures and utilizes the context relation by three modules: i) a Sequential Context Module captures consecutive relation with a GRU-like architecture, ii) an Interactive Context Module acquires cooperative context with global correlated linear fusion, and iii) a Biased Context Module benefits from the biased relation with distribution labels and the L1 loss. Extensive experiments on EMOTIC and CAER datasets show that our SIB-Net improves baseline significantly and achieves comparable results to the state-of-the-art methods.*

## 1. Introduction

Emotion clues from face, body, and scene play an important role in expressing messages and intentions. Context-aware emotion recognition aims to identify a person's emotion via all these clues, which has many potential applications including surveillance [6], games [19], educations [29], medical treatments [19] and marketing [20]. However, recognizing emotions from the context is a challenging problem.

In the past decade, researchers have made significant progress on human emotion recognition from visual contents [21]. They mainly focus on (1) facial expression analysis [12, 4, 26, 9, 30] and (2) body postures and gesture analysis [16, 13, 24]. Recently, some methods have made use of context information to boost the performance [3, 10, 11, 14], since faces or bodies are possibly indistinguishable and ambiguous in the wild. However, encoding effective emotion features from context information is still an open problem. Existing works mainly extract context information independently from the face, body and scene. These strategies may be limited in the understanding of complicated emotion context relations. According to the studies [15, 28, 17], current approaches fail to capture the sequential, interactive and biased context relation. The sequential context relation means that the current context's emotion is affected by the previous one in a head-body-scene order. The interactive context relation represents that all contexts' emotions depend on each other. The biased context relation indicates that different emotions occur in different contexts in one image.


Figure 1. Illustration of three issues. (a) Informationless or complicated contexts. The masking scene is too informationless to recognize emotions (b) Ambiguous contexts. Profiles, partial gestures and waves contain uncertain emotions. (c) Biased contexts. The sad agent plays games surrounded by sardonic partners.

Existing methods equally and independently deal with context information, which may lead to some problems illustrated in Figure 1. In Figure 1 (a), in informationless or complicated scenes, we feel difficult to identify emotions in the scene before considering the face and body. In Figure 1 (b), given ambiguous contexts like profiles, partial gestures and common scenarios, we feel less confident to decide emotions for each context. In Figure 1 (c), for biased contexts where different regions contain different emotions, it confuses the network's training with absolute labels.

*Corresponding author

In this paper, we address these complicated contexts with a Sequential Interactive Biased Network (SIB-Net). SIB-Net consists of three components: sequential context module (SCM), interactive context module (ICM), and biased context module (BCM). Firstly, we introduce SCM to capture sequential context relation, motivated by success of recurrent neural network family [5, 8] in sequential modeling. Specifically, we construct the recurrent feature to propagate the sequential influence in a face-body-scene order. To prevent the noise in the propagation process, we use a reset-gate and an update-gate to filter the invalid information. In this way, SIB-Net is capable to capture more semantic emotion information in context, even informationless or complicated scenes like Figure 1 (a).

Further, we present ICM to obtain the interactive context relation, inspired by the achievement of non-local [27] in capturing long-range dependencies. Intuitively, we may linearly fuse features from all context by several convolutional layers. It is expected that the fused feature contains interactive information of all contexts. To stress useful information, ICM considers a correlated linear fusion, where the fused features is multiplied with a correlated matrix. What is more, one branch not only considers individual interaction with other branches but also global interaction between all branches each other. Therefore, SIB-Net enhance semantic emotion information of each context, even all contexts are ambiguous like Figure 1 (b).

Additionally, we present BCM to leverage the biased context relation, stimulated by the advantage of Label Distribution Learning (LDL) [7] in dealing with label ambiguity. Given a one-hot label vector, we select random values from a Gaussian distribution to replace those zero values. The motivation is that there is no absolute emotion category after masking some context regions. To further boost performance, we replace the binary cross-entropy (bce) loss with the L1 loss, which has a tighter convergence bound for LDL classification. Consequently, as all context images has probability of all classes, they are effectively optimized even in biased contexts like Figure 1 (c).

Finally, we validate our method on public datasets. SIB-Net achieves promising quantitative and qualitative results, compared to state-of-the-art works and other fusion methods. Additionally, systematic ablation experiments show the rationality of every setting.

## 2. Related Work

**Context-Aware Emotion Recognition:** Recent works in context-aware emotion recognition are based on deep-learning network architectures. Kosti *et al*. [10] establish EMOTIC dataset and propose two-stream architectures followed by a fusion network. Lee *et al*. [11] build CAER dataset and mask the face from the image to capture semantic information in the context. Zhang *et al*. [31] use a Region Proposal Network to extract context elements from the image and a Graph Convolution Network to encode context. Mittal *et al*. [14] combine three interpretations of context (*e.g*. multiple modalities, semantic context and socio-dynamic interactions and proximity) for emotion recognition. All these works mainly extract context representations of the face, body and scene independently.

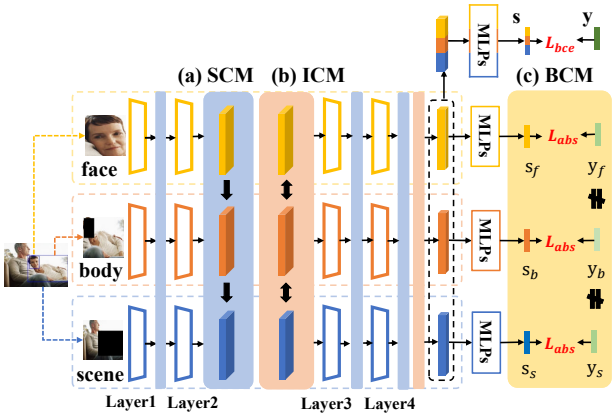


Figure 2. The pipeline of SIB-Net. It is built on three branches including face (deep yellow), body (deep orange) and scene (deep blue), and three inserted modules including SCM (light blue), ICM (light orange) and BCM (light yellow).

**Emotional Context Relation:** Nardone [15] designs experiments to validate that the intensity and trajectory of emotions are influenced by the sequence in which they are experienced. Wieser *et al*. [28] demonstrate that perception and neural processing of facial expressions are substantially modified by contextual information. Panda *et al*. [17] identify an important but overlooked issue that there is emotion bias between visual benchmark datasets. The above studies show that context emotion relation contains the sequential, interactive and biased context relation.

## 3. SIB-Net

To exploit context emotion information of the face, body and scene, we propose a simple yet effective framework SIB-Net for context-aware emotion recognition. In this section, we first describe an overview of SIB-Net and then introduce its three modules.

### 3.1. Overview

SIB-Net is built upon traditional CNNs, as shown in Figure 2. Given an image, we extract its face, body and scene into masking partial images. These images are fed into the three-branch backbone for feature extraction. For training and testing, all features are concatenated and classified by Multiple Linear Perceptions (MLPs). In basic setting, we only train the network with the bce loss ($L_{bce}$) between the predictive score $s$ and the ground-truth label $y$.

To get the relation information between three branches, SIB-Net introduces three modules: (a) SCM adopts GRU-like architecture to obtain the sequential relation feature. We make all layers' features as SCM's input and sum them with SCM's output. (b) ICM introduces global correlated linear fusion to calculate the interactive relation feature. We make the 2nd and 4th layers' features as ICM's input and sum them with ICM's output. (c) BCM leverages label distribution learning and the L1 loss ($L_{abs}$) to assist the optimization of the face, body, and scene branches. We further train three branches with the L1 loss between their scores ($s_f$, $s_b$ and $s_s$) and the distribution labels ($y_f$, $y_b$ and $y_s$).

### 3.2. Sequential Context Module

The sequential context relation is crucial for context-aware emotion recognition, e.g. , emotions of the different regions interacted each other in a face-body-scene order. Existing works ignore this relation possibly because of the difficulty of modeling the sequential cognitive mechanism. The RNN family [5, 8] succeeds in using the recurrent connection and the gated mechanism for sequence modeling. Motivated by it, we apply GRU-like architecture to design SCM for processing emotional features orderly.

We use the recurrent features to propagate the sequential influence from the face branch, the body branch to the scene branch, which can be formulated as:

$$o_t, h_t = f\left(x_t, h_{t-1}\right) \tag{1}$$

where $t = f, b, s$ denotes the branch of face, body and scene respectively, $x$, $h$ and $o$ denotes input features, recurrent features and output features, $f$ denotes the mapping function. In experiments, $x$ is obtained by global average pooling of the layer's output feature, and $o$ is expanded as the original size and added with the layer's output feature. It is expected that the current emotion is shaped by the current emotional feature and the former emotion. We describe the mapping function $f$ in the following content.

To prevent the noise propagating from the former branch to the current one, we use a reset gate $R_t$ to filter invalid information. As shown in Eq. 2, we concatenate $h_{t-1}$ and $x_t$, transform it linearly with a learnable matrix $W_r$ and bias $b_r$, and then use sigmoid function $\sigma$ to normalize its value into $[0, 1]$. We denote this value as $r_t$ and assign it to $h_{t-1}$ by element-wise multiplication. Intuitively, $R_t$ controls the influence of $h_{t-1}$. If $x_t$ expresses the totally different emotion from $h_{t-1}$, $R_t$ outputs zero to ignore $h_{t-1}$. Then, we concatenate $h_{t-1} \otimes r_t$ and $x_t$, transform it linearly with a learnable matrix $W_h$ and bias $b_h$, and then use a hyperbolic tangent function function $tanh$ to get $\hat{h}_t$. $\hat{h}_t$ is supposed to be the primitive hidden emotion for $t$-th branch.

For suppressing the noise that the current emotion propagates to the whole emotion, we adopt an update gate $U_t$ to eliminate detrimental information. As shown in Eq. 2, we

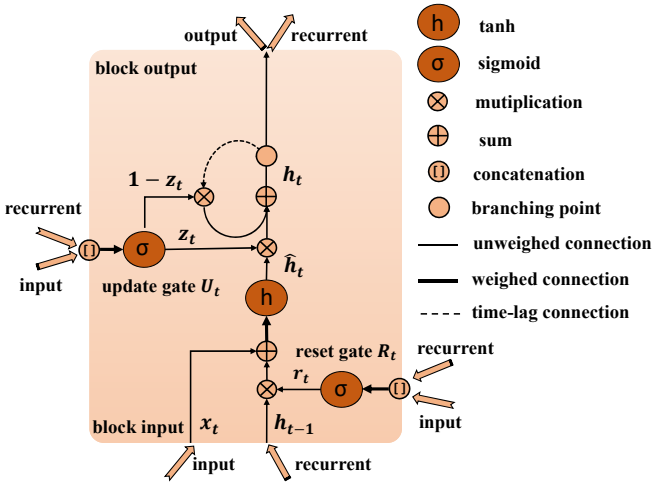


Figure 3. Configuration of Sequential Context Module.

concatenate $h_{t-1}$ and $x_t$, transform it linearly with a learnable matrix $W_z$ and bias $b_z$, and then use sigmoid function $\sigma$ to normalize its value into $[0, 1]$. We denote this value as $z_t$. Then we assign $z_t$ to $\hat{h}_t$ and $1 - z_t$ to $h_{t-1}$ by element-wise multiplication, and get $h_t$ by summation. Through such a smoothing method, we get a more effective hidden emotion.

$$\begin{aligned} r_t &= \sigma\left(W_r\left[h_{t-1}, x_t\right] \oplus b_r\right) \\ z_t &= \sigma\left(W_z\left[h_{t-1}, x_t\right] \oplus b_z\right) \\ \hat{h}_t &= tanh\left(W_h\left[h_{t-1} \otimes r_t, x_t\right] \oplus b_h\right) \\ h_t &= \left(1 - z_t\right) \otimes h_{t-1} \oplus z_t \otimes \hat{h}_t \end{aligned} \tag{2}$$

We can understand SCM more clearly from Figure 3. Specifically, we insert one SCM after each layer of the backbone. It is the first work to model the emotional sequential relation in context. Compared to the vanilla RNN and LSTM [8], GRU is more suitable and benefits from two advantages: 1) It prevents noise in the sequential influence, 2) It contains fewer parameters to optimize.

### 3.3. Interactive Context Module

The interactive context relation is important for context-aware emotion recognition, e.g. , humans identify the image's emotion by jointly considering partial images. However, current methods fail to capture this relation because computing an effective instead of detrimental interactive context relation is a nontrivial problem. Non-local [27] achieves good performance using non-local operations in capturing long-range dependencies. Inspired by it, we propose ICM to obtain the multiple weighted interactive information between face, body and scene.

Given a feature from one branch with size $C$, $H$ and $W$ for channel, height and width, we concatenate three features along channel to get the tensor $X$. We process $X$ with two convolutional layers to get the tensor $Y$. It is expected that $Y$ contains interactive information of the face, body and

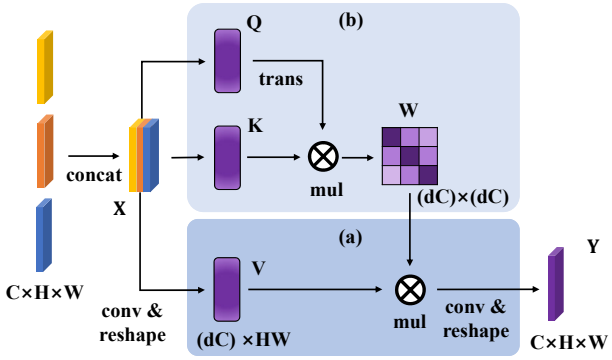

Figure 4. Configuration of Interactive Context Module.

scene, as $Y$ is the linear combination of their features. Usually, we use a factor $d$ to reduce inner features' channel.

To enhance useful interactive information, ICM considers a correlated linear fusion with a learnable correlation matrix. As shown in Figure 4, we process $X$ with different single convolutional layers and reshape operation to get $\{Q, K, V\} \in \mathbb{R}^{dC \times HW}$. Then we transpose $Q$ and multiply it with $K$ to get a matrix $W \in \mathbb{R}^{dC \times dC}$. Each element in $W$ indicates the similarity between two features. Higher value can be interpreted as more effective interaction. For numerical stability, we divide $W$ with the square of the 2nd dimension, and use a softmax function to normalize $W$ along the 2nd dimension. Then, we use $W$ to multiply $V$ so that all features are correlated linear fused. Finally, we process the fused features to get $Y \in \mathbb{R}^{C \times H \times W}$ with one convolutional layer and reshape operation. This strategy makes the interactive information more qualified.

Unlike Non-local [27], which takes one block for all branches, we assign each branch with different ICM. Therefore, one branch not only considers individual interaction between it and other branches but also global interaction between all branches each other. We find global interaction is richer and more effective experimentally. In experiments, we place individual SCM after the 2nd and 4th layers.

### 3.4. Biased Context Module

The biased context relation should be noticeable in the context-aware emotion recognition, e.g. , an emotional gap possibly occurs between the whole image and the masking partial images. Present study [17] tries to eliminate this bias by introducing additional data instead of making use of it. On the other hand, LDL [7] utilizes the label inconsistency and relativity between classes to boost the performance. Stimulated by it, we present BCM to leverage label distribution learning to optimize the face, body, and scene branches.

We generate discrete distribution labels $y_f$, $y_b$ and $y_s$ for the face, body and scene branch. Specifically, given a ground-truth one-hot label vector $y$, we select random values from a Gaussian distribution to replace those zero values. Since there is no absolute emotion category after masking some context regions. In experiments, we set the variance to 0.1 and the mean to 0.4 for Gaussian distribution. In this way, all context images are effectively used for optimizing, as every image has the probability of all classes. Through this technique, we model the biased relation between the whole image and the masking partial images, and make use of it to enhance performance. Compared to traditional LDL, we do not require the sum of label vector should be 1 for the following reasons: 1) We are motivated by the label inconsistency from masking instead of label relation. 2) Traditional LDL sets the label to an too small value to damage to the network's optimization.

Further, we replace the bce loss with the L1 loss. According to Wang [25], the L1 loss is a better measure in applying LDL for classification as it has a tighter convergence bound. As shown in Figure 2, we train the network with the bce loss $L_{bce}$ for the whole image and the L1 loss $L_{abs}$ for the masking partial images.

## 4. Experiments

### 4.1. Datasets

**EMOTIC** [10] contains 23,571 in-the-wild images and 34,320 annotated people. The annotators assign 26 categorical and VAD dimensional emotions for each image. These images are divided into three sets: training images (70%), validation images (10%) and testing images (20%). We use 26 categorical emotions as annotations and mean Average Precision (mAP) for the evaluation metric.

**CAER-S** [11] includes 70,000 in-the-movie images annotated with 6 basic categorical emotions and a neural emotion. These images are randomly split into training (70%), validation (10%) and testing (20%) images. We use the Overall Sample Accuracy (OSA) for measurement.

### 4.2. Implementation Details

**Backbone:** We choose three ResNet-18 as the backbone like previous works. By default, the ResNet-18 is pretrained on the ImageNet [22] for better initialization. The emotional features are extracted from the last pooling layers and concatenated together for classification.

**Pre-processing:** We use the face's and body's bounding box to get the masking image for each branch. Given the bounding box $bbox$, the mask image $I_{mask}$ is given as:

$$I_{mask} = \begin{cases} I(i,j) & I(i,j) \notin bbox \\ 0 & otherwise \end{cases} \quad (3)$$

For EMOTIC [10], we use the provided bounding boxes of the face and the body. While for CAER [11], we use OpenFace [1] and OpenPose [2] to detect the bounding boxes.

| Method | Tools | | | EMOTIC (mAP) | CAER-S (OSA) |
|---|---|---|---|---|---|
| | Fa | Bo | De | | |
| Kosti *et al*. [10] | + | + | - | 27.38% | - |
| Lee *et al*. [11] | + | + | - | 33.13% | 73.51% |
| Mittal *et al*. [14] | + | + | + | **35.48**% | - |
| concatenation | + | + | - | 33.33% | - |
| attention | + | + | - | 33.13% | - |
| SIB-Net | + | + | - | **35.41**% | **74.56**% |

Table 1. **Quantitative results**. We show quantitative results and auxiliary tools usage of existing works and SIB-Net. *Fa*, *Bo* and *De* means following auxiliary tools: face detection, body detection and depth detection. + indicates used while - means unused. SIB-Net has promising accuracy and uses few auxiliary tools.

The mask images are resized to $230 \times 230$ and their values are normalized into $[-1, 1]$.

**Training:** We implement our SIB-Net with PyTorch [18] and train it with Nvidia Titan V. We set a batch size of 32 for EMOTIC and a batch size of 64 for CAER-S. SIB-Net is trained for 10 epochs for EMOTIC and 20 epochs for CAER-S with an Adam optimizer. The initial learning rate of the backbone's parameters is set as 0.0001. While the initial learning rate of the classifier's and the modules' parameters is set as 0.001. During the training, the learning rate is multiplied with 0.1 at the 0.5, 0.75 and 0.875 phase of the whole epochs. The whole network is jointly optimized with $L_{bce}$ and $L_{abs}$ with the ratio 1:1. To avoid the influence of random factors, we average the results of three experiments as the final score.

### 4.3. Quantitative and Qualitative Comparision

We compare SIB-Net with other methods quantitatively and qualitatively. Kosti *et al*. [10] and Mittal *et al*. [14] adopt concatenation fusion for context features. Lee *et al*. [11] achieve the state-of-the-art (sota) accuracy and leverage attention mechanism for fusion. We utilize concatenation fusion in baseline method. SIB-Net can be regarded as sequential interactive biased fusion.

We show quantitative results and auxiliary tools usage of existing works and SIB-Net. As shown in Table 1, SIB-Net achieves promising accuracy with 35.41% on EMOTIC and 74.56% on CAER-S. It is worth to mention that SIB-Net needs fewer tools than the sota method, which may accounts for relative lower accuracy.

Additionally, SIB-Net performs better than attention fusion by 2.28% on EMOTIC. We also show Class Activation Maps (CAM) [23] on features outputed from the networks' 4th layer. As shown in Figure 5, SIB-Net captures more semantic regions not only in normal scene but also in the ambiguous, informationless or complicated context. This is because SIB-Net considers the relation of different contexts.

### 4.4. Ablation Studies

We show ablation studies in Table 2 to demonstrate rationality of every setting. Several observations can be concluded in the following.

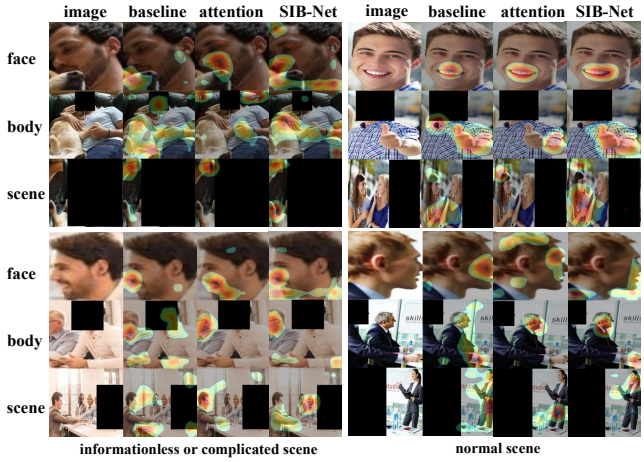


Figure 5. **Qualitative results**. We show CAM [23] of different fusion methods. The red means the region the model pays attention to. SIB-Net captures more semantic regions not only in normal scene but also in informationless and complicated context.

Compared to using single context like face, body and scene, combining all context information increases accuracy at least 0.67% on EMOTIC. We denote the context combination as baseline.

For sequential modeling, GRU-like architecture achieves better accuracy than vanilla RNN and LSTM. It suggests that 1) GRU-like architecture prevents noise in the sequential influence compared to vanilla RNN architecture, 2) GRU-like architecture contains fewer parameters to optimize compared to LSTM architecture.

For interactive modeling, global correlated linear fusion (gclf) achieves better performance than linear fusion (lf) and correlated linear fusion (clf). It indicates that 1) the clf effectively suppresses the detrimental influence through correlation weighting mechanism compared to the lf, 2) the gclf contains richer information compared to the clf.

For biased modeling, our modified LDL (LDL*) with the L1 loss achieves better results than LDL and LDL*. It claims that 1) LDL* is more effective for biased modeling as the probability values are more reasonable, 2) the L1 loss is a better measure for applying LDL for classification.

Capturing every relation of context is significantly important to emotion recognition. Specifically, ICM is more effective than SCM and BCM. It indicates that the interactive relation is most important in emotion perception. What is more, the effectiveness increases sub-linearly as more modules are added. It indicates these relations are complemented and intersected.

## 5. Conclusion

This paper proposes a Sequential Interactive Biased Network (SIB-Net) to capture the context relation for context-aware emotion recognition. We demonstrate the ef-

| Method | EMOTIC (mAP) |
|---|---|
| face | 32.66% |
| body | 29.34% |
| scene | 29.57% |
| baseline | **33.33**% |
| baseline+S (*vanilla RNN*) | 34.17% |
| baseline+S (*LSTM*) | 33.99% |
| baseline+S (*GRU*) | **34.21**% |
| baseline+I (*lf*) | 33.71% |
| baseline+I (*clf*) | 34.09% |
| baseline+I (*gclf*) | **34.33**% |
| baseline+B (*LDL*) | 33.58% |
| baseline+B (*LDL**) | 34.02% |
| baseline+B (*LDL*+L1*) | **34.12**% |
| baseline+S (*GRU*)+I (*gclf*) | 35.14% |
| baseline+S (*GRU*)+B (*LDL*+L1*) | 34.57% |
| baseline+S (*GRU*)+I (*gclf*)+B (*LDL*+L1*) | **35.41**% |

Table 2. **Ablation Studies**. S, I and B denote sequential modeling, interactive modeling and biased modeling. The content in brackets denotes different modeling methods.

fectiveness of SIB-Net through quantitative and qualitative experiments. SIB-Net achieves promising accuracy with 35.41% on EMOTIC and 74.56% on CAER-S. Besides, SIB-Net captures more semantic regions not only in normal scene but also in the ambiguous, informationless or complicated context. This is because SIB-Net considers the relation of different contexts through three modules: SCM, ICM and BCM. It is the first work to capture the complicated context relations and indicate the importance of them in context-aware emotion recognition. In the future, we will extend our work to more datasets including video. What is more, due to the complexity of relations, it still needs more efforts to find a simplified and effective method to capture the relations.

**Acknowledgements:** This work is supported by the Stable Support Projects for Shenzhen Higher Education Institutions (SZWD2021011), the Scientific Research Platforms and Projects in Universities in Guangdong Province (2019KTSCX204), National Natural Science Foundation of China under Grant (61702193), and the Fundamental Research Funds for the Central Universities of China under Grant (2019JQ01). Thanks for the helpful advice from Xin Lin, Xubin Zhong, Jiapeng Tang, Tong Zhou and Wei Liu.

## References

[1] T. Baltrusaitis, P. Robinson, and L. P. Morency. Openface: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 4

[2] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 4

[3] C. Chen, Z. Wu, and Y.-G. Jiang. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *ACM MM*, pages 127–131, 2016. 1

[4] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, pages 13984–13993, 2020. 1

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2, 3

[6] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008. 1

[7] B. B. Gao, C. Xing, C. W. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *TIP*, PP(99):1–1, 2016. 2, 4

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 3

[9] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *CVPR*, pages 9841–9850, 2019. 1

[10] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Context based emotion recognition using emotic dataset. *TPAMI*, 42(11):2755–2766, 2019. 1, 2, 4, 5

[11] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn. Context-aware emotion recognition networks. In *CVPR*, pages 10143–10152, 2019. 1, 2, 4, 5

[12] S. Li and W. Deng. Deep facial expression recognition: A survey. *TAC*, 2020. 1

[13] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *IJCV*, 128(1):1–25, 2020. 1

[14] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *CVPR*, pages 14234–14243, 2020. 1, 2, 5

[15] S. Nardone. Are emotions influenced by their sequence? an experimental study of emotional processing. 2019. 1, 2

[16] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari. Survey on emotional body gesture recognition. *TAC*, 2018. 1

[17] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *ECCV*, pages 579–595, 2018. 1, 2, 4

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 5

[19] G. Pioggia, R. Igliozzi, M. Ferro, A. Ahluwalia, F. Muratori, and D. De Rossi. An android for enhancing social skills and emotion recognition in people with autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(4):507–515, 2005. 1

[20] F. Ren and C. Quan. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, 13(4):321–332, 2012. 1

[21] P. V. Rouast, M. Adam, and R. Chiong. Deep learning for human affect recognition: Insights and new developments. *TAC*, 2019. 1

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 4

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 5

[24] Z. Shen, J. Cheng, X. Hu, and Q. Dong. Emotion recognition based on multi-view body gestures. In *ICIP*, pages 3317–3321. IEEE, 2019. 1

[25] J. Wang and X. Geng. Classification with label distribution learning. In *IJCAI*, pages 3712–3718, 2019. 4

[26] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, pages 6897–6906, 2020. 1

[27] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2, 3, 4

[28] M. Wieser and T. Brosch. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, 3:471, 2012. 1, 2

[29] D. Yang, A. Alsadoon, P. C. Prasad, A. K. Singh, and A. Elchouemi. An emotion recognition model based on facial recognition in virtual learning environment. *Procedia Computer Science*, 125:2–10, 2018. 1

[30] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *CVPR*, pages 2168–2177, 2018. 1

[31] M. Zhang, Y. Liang, and H. Ma. Context-aware affective graph reasoning for emotion recognition. In *ICME*, 2019. 2