Prototype Learning for Micro-gesture Classification

Guoliang Chen[1], Fei Wang[1], Kun Li[4,*], Zhiliang Wu[4], Hehe Fan[4], Yi Yang[4], Meng Wang[1,2,3] and Dan Guo[1,2,3,5,*]

[1]*School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology (HFUT)*

[2]*Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education*

[3]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China*

[4]*CCAI, Zhejiang University, China*

[5]*Anhui Zhonghuitong Technology Co., Ltd.*

Abstract

In this paper, we briefly introduce the solution developed by our team, HFUT-VUT, for the track of Micro-gesture Classification in the MiGA challenge at IJCAI 2024. The task of micro-gesture classification task involves recognizing the category of a given video clip, which focuses on more fine-grained and subtle body movements compared to typical action recognition tasks. Given the inherent complexity of micro-gesture recognition, which includes large intra-class variability and minimal inter-class differences, we utilize two innovative modules, *i.e.*, the cross-modal fusion module and prototypical refinement module, to improve the discriminative ability of MG features, thereby improving the classification accuracy. Our solution achieved significant success, **ranking 1st** in the track of Micro-gesture Classification. We surpassed the performance of last year's leading team by a substantial margin, improving Top-1 accuracy by 6.13%.

Keywords

Micro-gesture, action classification, multi-model action recognition, video understanding

1. Introduction

Micro-gestures (MGs) [1, 2] are defined as a special category of body gestures that are indicative of humans' emotional status. Examples include "head scratching", "nose rubbing" and "hand rubbing", which are not intended to be communicative but arise spontaneously from stress or discomfort. Unlike indicative gestures, which are intended to facilitate communication, micro-gestures have not been well studied, and existing works are focused primarily on macro-body movements, neglecting these subtle gestures and their connection to hidden emotions.
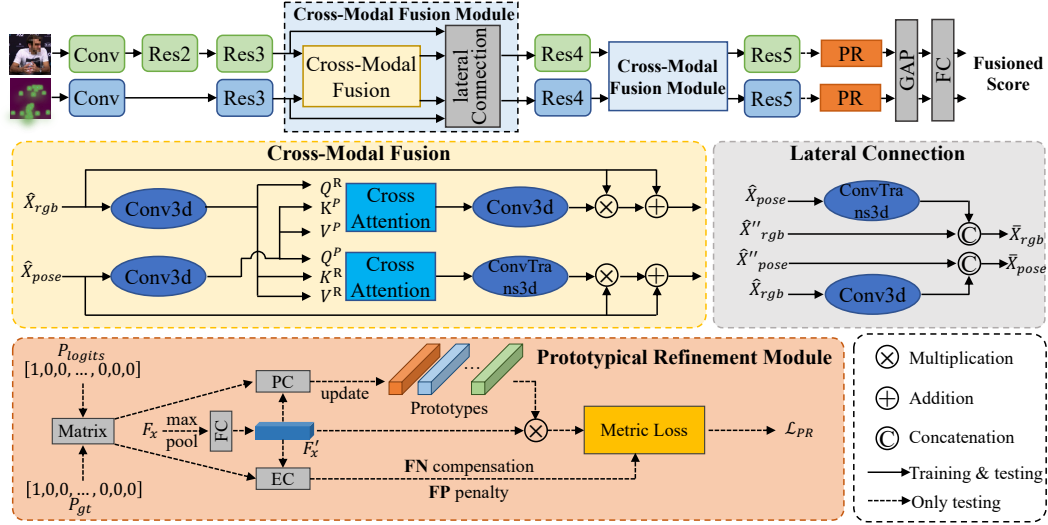
Figure 1: Overview of the proposed method for micro-gesture classification. The network contains the RGB and Pose branches. The cross-attention fusion module is used to explore the correlations between RGB and Pose modalities. The prototypical refinement module defines prototype representations of each fine-grained micro-gesture category during training and forces the model to calibrate ambiguous samples among different micro-gesture categories.

Compared to ordinary action or gesture recognition, MGs present more challenges. MGs encompass more refined and subtle bodily movements that occur spontaneously during real-life interactions. Additionally, micro-gestures suffer from small inter-class differences and large intra-class differences. Therefore, micro-gestures that have unclear action boundaries, large individual differences, high similarity, and are easily influenced by the environment tend to be misclassified, which we refer to as ambiguous samples.

In this challenge, we adopt PoseConv3D [3] as the baseline model, and the main contributions of our method are summarized as follows:

- We proposed a multi-modality-based network for micro-gesture classification. Specifically, we incorporate the cross-modal fusion module and prototypical refinement module for action classification.
- For the micro-gesture classification challenge, our method achieves a Top-1 accuracy of 70.254 on the iMiGUE test set. The experimental results indicate that our method effectively captures subtle changes in micro-gestures.

2. Methodology

2.1. Network Architecture

The main structure of the proposed method is illustrated in Figure 1. We adopt the PoseConv3D [3] network as the backbone, which enables more efficient learning of spatio-temporal features and greater robustness in noise handling. Concretely, the proposed method consists

of a two-pathway 3D CNN-based backbone network, where the upper path is responsible for processing RGB data, and the lower path processes skeleton data.

The inputs to the model are $X_{rgb} = (x_1, ..., x_{t_{rgb}})$ and $X_{pose} = (x_1, ..., x_{t_{pose}})$, where t_{rgb} and t_{pose} represent the number of frames, we set t_{rgb} to 8 and t_{pose} to 32. The backbone network comprises the Cross-Modal Fusion Module and Prototypical Refinement Module. Specifically, the Cross-Modal Fusion Module aims to explore the correlations between different modalities. The Prototypical Refinement Module defines prototype representations of each fine-grained micro-gesture category during training and forces the model to calibrate ambiguous samples among different micro-gesture categories. These feature vectors are subsequently mapped to the probability distribution of K candidate categories through a fully connected (FC) layer with softmax activation. Finally, the probability distributions of the two modalities are fused in a 1:1 ratio to obtain the model's final prediction for the micro-gesture category.

2.2. Cross-Modal Fusion Module

Due to the subtle and complex nature of micro-gestures, different categories of micro-gestures may have different advantages and disadvantages in terms of expression in different modalities. To effectively leverage the characteristic features of different modalities within the backbone network, we propose the Cross-Modal Fusion module.

Assuming the features of the module are $\hat{X}_{rgb} \in \mathbb{R}^{T_1 \times H_1 \times W_1 \times C_1}$ and $\hat{X}_{pose} \in \mathbb{R}^{T_2 \times H_2 \times W_2 \times C_2}$. Firstly, these features are mapped to the same hidden dimension through two 3D convolutional layers, and then a maximum pooling layer is used to compress the spatial features to (1,1). At this point, $\hat{X}_{rgb}, \hat{X}_{pose} \in (T_1, C')$, where C' represents the hidden dimension. Subsequently, we utilize the cross-attention fusion mechanism to fuse the features learned from RGB and Skeleton modalities in a multi-level adaptive manner. We define the cross-attention fusion mechanism as:

$$\text{Attn}_m \in F_{attn}(Q^m, K^m, V^m), \text{where } m \in \{\text{RGB, Pose}\}. \tag{1}$$

The variables in Equation 1 are defined as follows:

$$Q^m = W^{q_m}\hat{X}'_m, K^m = W^{k_m}\hat{X}'_m, V^m = W^{v_m}\hat{X}'_m. \tag{2}$$

In contrast to existing methodologies that implement attention [4, 5, 6] and cross-attention mechanisms [7, 8, 9, 10] along spatial dimensions, this paper introduces a novel approach that applies cross-attention based on channel dimensions. This technique enables the model to execute information fusion across different modalities, specifically at the channel level, and facilitates the learning of correlations between channel-dimensional information across these modalities. Additionally, it effectively minimizes redundant and noisy data within the channels. By focusing cross-attention on channel dimensions rather than the entire feature map, this method significantly reduces computational complexity, offering a more efficient solution for processing multimodal data.

Finally, the resulting attentional weights are multiplied with the inputs in the channel dimension and then summed:

$$\begin{cases} \hat{X}''_{rgb} = (Conv3d(\hat{X}'_{rgb}) \cdot \hat{X}_{rgb}) + \hat{X}_{rgb}, \\ \hat{X}''_{pose} = (ConvTrans3d(\hat{X}'_{pose}) \cdot \hat{X}_{pose}) + \hat{X}_{pose}. \end{cases} \tag{3}$$

Finally, the fused $\hat{X}''_{rgb}$ and $\hat{X}''_{pose}$ are spliced through lateral connections:

$$
\begin{cases}
\bar{X}_{rgb} = [Conv3d(\hat{X}_{pose}); \hat{X}''_{rgb}], \\
\bar{X}_{pose} = [ConvTrans3d(\hat{X}_{rgb}); \hat{X}''_{pose}],
\end{cases}
\tag{4}
$$

where[;] denotes the concatenation of two features.

2.3. Prototypical Refinement Module

Micro-gesture recognition suffers from the problem of large intra-class differences and small inter-class differences. In order to solve this problem, we utilize the prototypical refinement module inspired by [11], which uses prototype-based contrastive learning, and is dedicated to discovering and calibrating the ambiguous samples between these categories. As shown in Figure 1, we firstly utilize backbone network to extract video feature as $F_x \in \mathbb{R}^{N \times C \times H \times W}$, and preliminary categorical predictions are calculated as $P_{logits} \in \mathbb{R}^{N \times K}$, where K represents the number of action categories for micro-gestures. The ground-truth is also obtained as $P_{gt} \in \mathbb{R}^{N \times K}$. The extracted feature F_x is then fed into a maximum pooling layer and a fully connected layer, and then we can get $F'_x \in \mathbb{R}^{N \times D}$, where D is the hidden dimension.

Ambiguous Samples Discovery. Given an action label k, using P_{logits} and P_{gt}, we can find the confidence samples and ambiguous samples in the training phase as follows. If a sample is predicted correctly, namely as a True Positive (TP), we consider it a confident sample to distinguish it from ambiguous samples. If a sample of action k is misclassified as another category, it is called a False Negative (FN). If samples from other categories are misclassified as action k, it is called False Positive (FP). We find these ambiguous samples in a batch and compute their center representations as follows:

$$
\mu_{FN}^k = \frac{1}{n_{FN}^k} \sum_{j \in s_{FN}^k} F_j, \quad \mu_{FP}^k = \frac{1}{n_{FP}^k} \sum_{j \in s_{FP}^k} F_j.
\tag{5}
$$

Prototype Clustering. We created a prototype representation for each micro-gesture category and randomly initialized it before training. Take the action category k as an example, the prototypes $\mathcal{P}_k$ serve as a stable estimate of the clustering center for action category k and can be continuously optimized during the training process through confidence samples (namely true positive TP samples) of category k. With online training, the prototype $\mathcal{P}_k$ of action category k can be updated by exponential moving average (EMA):

$$
\mathcal{P}_k = (1 - \rho) \cdot \frac{1}{n_{TP}^k} \sum_{i \in s_{TP}^k} F_i + \rho \cdot \mathcal{P}_k^{pre},
\tag{6}
$$

where F_i is the feature of sample i, $\mathcal{P}_k^{pre}$ is the prototype before updating. ρ is a momentum term and is empirically set to 0.9.

Error Calibration. To calibrate the prediction of ambiguous samples, we use the confidence sample i of action k as an anchor and compute the two auxiliary terms in the feature space, which are ϕ_i and φ_i for FN and FP samples, respectively, and are defined as follows:

$$
\phi_i = \begin{cases}
1 - dis(F_i, \mu_{FN}^k), if\, i \in s_{TP}^k \, and \, n_{FN}^k > 0 \\
0, \, otherwize
\end{cases},
\tag{7}
$$

$$\varphi_i = \begin{cases} 1 + dis(F_i, \mu_{FP}^k), \, if \, i \in s_{TP}^k \, and \, n_{FP}^k > 0 \\ \qquad 0, \; otherwize \end{cases}, \tag{8}$$

where $dis(,)$ denotes the cosine distance between two features. ϕ_i, as the compensation term for the set of FN samples, should be closer to the confidence samples in the feature space by minimizing ϕ_i, motivating the model to correct the ambiguous samples as action k. φ_i, as the penalty term for the FP samples, should be further away from the confidence samples in the feature space by minimizing φ_i, preventing the model from identifying the ambiguous samples as action k. Finally, its prototypical refinement loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{PR}(i) = &- \log \frac{dis(F_i,P_k)/\tau - (1-p_{ik})\varphi_i}{dis(F_i,P_k)/\tau - (1-p_{ik})\varphi_i + \sum_{l \neq k} dis(F_i,P_l)/\tau}, \\ &- \log \frac{dis(F_i,P_k)/\tau - (1-p_{ik})\phi_i}{dis(F_i,P_k)/\tau - (1-p_{ik})\phi_i + \sum_{l \neq k} dis(F_i,P_l)/\tau}, \end{aligned} \tag{9}$$

where p_{ik} is the predicted probability score of sample i for the micro-gesture category k.

Finally, the total loss in the training process can be defined as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{PR}, \tag{10}$$

where $\mathcal{L}_{CE}$ is Cross-Entropy loss and α is a hyper-parameter.

3. Experiments

3.1. Datasets

iMiGUE [1] dataset. This dataset comprises 32 micro-gestures, along with one non-micro-gesture class, collected from post-match press conference videos of tennis players. This challenge follows a cross-subject evaluation protocol, wherein the 72 subjects are divided into a training set consisting of 37 subjects and a testing set comprising 35 subjects. For the MG classification track, 12,893, 777, and 4,562 MG clips from iMiGUE are used for train, val, and test, respectively. In this challenge, it is allowed to use RGB and skeleton modal data.

3.2. Evaluation Metrics and Implementation Details

For the micro-gesture classification challenge, we calculate the Top-1 Accuracy to assess the prediction results. MMAction2 [12] toolbox has been used for the implementation of our approach. Since we use PoseConv3D [3] as the baseline method, we first need to train the RGB and Pose branches individually and fuse the best weights obtained from these two modalities as the initial weights for the two-branch method. The frame sizes for RGB and poses are empirically set to 8 and 32, respectively. The Stochastic Gradient Descent (SGD) optimizer is employed with a momentum of 0.9 and a weight decay of 1e-4 in the training process. We set the batch size to 10 and the initial learning rate to 0.0075. The learning rate is reduced by a factor of 10 at the 8-th and 22-th epochs, and the model is trained with 30 epochs.

Table 1

The top-3 results of micro-gesture classification on the iMiGUE test set. Data is provided by the Kaggle competition page[1].

Rank	Team	Top-1 Accuracy (%)
1	HFUT-VUT (**Ours**)	**70.254**
2	NPU-MUCIS	70.188
3	ywww11	68.917

Table 2

The results of micro-gesture classification on the test set of the iMiGUE dataset. J and L denote the joint and limb skeleton data, respectively.

Method	Modality	Top-1 Accuracy (%)
TSM [13]	RGB	58.77
VSwin-T [14]	RGB	59.97
VSwin-S [14]	RGB	57.83
VSwin-B [14]	RGB	61.73
ST-GCN [15]	Skeleton (J)	46.38
ST-GCN++ [16]	Skeleton (J)	49.56
StrongAug [16]	Skeleton (J)	53.13
AAGCN [17]	Skeleton (J)	54.73
CTR-GCN [18]	Skeleton (J)	53.02
DG-STGCN [19]	Skeleton (J)	49.56
PoseConv3D [3]	Skeleton (J)	61.11
Li *et al.* (MiGA'23 1st) [20]	Skeleton ($J + L$)	64.12
EHCT (MiGA'23 2nd) [21]	Skeleton (L)	63.02
Ours (PoseConv3D)	Skeleton (J)	67.91
Ours (Ensemble Model)	RGB & Skeleton ($J + L$)	**70.25**

3.3. Experimental Results

As shown in Table 1, we first report top-3 results on the test set of the iMiGUE dataset. Our team achieves the best Top-1 Accuracy of 70.254. Compared to the team of "ywww11", our method outperforms it by 1.94%. In addition, we also compare our approach with skeleton-based and RGB-based action recognition methods on the iMiGUE datasets.

As shown in Table 2, we explore the action recognition methods using different modalities on the test set of the iMiGUE dataset. For the RGB-based methods, compared to TSM [13], Video Swin Transformer [14] is able to capture the local feature information better by local self-attention operation and achieves an excellent Top1-Accuracy result of 61.73. For the skeleton-based method, Li *et al.* [20] incorporates a semantic embedding loss to improve action classification performance and achieves the Top-1 accuracy of 64.12. Huang *et al.* [21] propose an ensemble hypergraph-convolution Transformer equipped with an auxiliary classifier to mitigate the impact of imbalanced data. In contrast, our method utilizes a multi-modal fusion strategy based on RGB and skeleton modalities to bridge the information differences between

---

[1]The Kaggle competition page: https://www.kaggle.com/competitions/2nd-miga-ijcai-challenge-track1/leaderboard

modalities and capture the information of micro-gestures more comprehensively. Since there are very similar features between different micro-gesture categories, the prototypical refinement module is used to calibrate these ambiguous micro-gesture samples. As a result, our method achieves 67.91 on the Top-1 accuracy. To incorporate the advantages of RGB-based and skeleton-based methods, we perform a model ensemble on the results of the Video Swin Transformer [14] base version and Li *et al.* [20] and our method. Finally, our method achieves 70.25 on the Top-1 accuracy, which is 6.13% higher than the results of MiGA'23 1st place.

4. Conclusion

In this paper, we present an innovative approach based on prototype learning for micro-gesture recognition in the MiGA challenge hosted at IJCAI 2024. Our approach adopts the PoseConv3D [3] model as the baseline, incorporating the cross-modal fusion module and prototypical refinement module for action classification. The RGB modality contains rich texture information, and the pose modality provides geometric information about the human structure. Through the cross-modal fusion module, the model can establish a correlation between the two modalities and combine the information of both modalities. The prototypical refinement module can provide a stable and clear representation of each category by creating a prototype for each micro-gesture action category. This module effectively mitigates intra-category differences in micro-gestures and makes the distinction between categories clearer. Eventually, the proposed approach reached 70.254 on the test set of the iMiGUE dataset.

In future work, we will address the issues in this challenge from other perspectives. One promising direction is the application of video motion magnification techniques [22, 23] to magnify the subtle changes of micro-gestures for better recognition. Additionally, we plan to pre-train our model on the large-scale micro-action dataset [24, 25] to investigate the impact of prior knowledge for micro-gesture recognition. We also plan to exploit the weakly supervised method [26] or unsupervised method [27] to alleviate the difficulty of data annotation in micro-gesture recognition.

Acknowledgments

References

[1] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10631–10642.

[2] H. Chen, H. Shi, X. Liu, X. Li, G. Zhao, Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis, International Journal of Computer Vision 131 (2023) 1346–1366.

[3] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.

[4] K. Li, J. Li, D. Guo, X. Yang, M. Wang, Transformer-based visual grounding with cross-modality interaction, ACM Transactions on Multimedia Computing, Communications and Applications 19 (2023) 1–19.

[5] S. Tang, R. Hong, D. Guo, M. Wang, Gloss semantic-enhanced network with online back-translation for sign language production, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5630–5638.

[6] K. Li, D. Guo, M. Wang, Vigt: proposal-free video grounding with a learnable token in the transformer, Science China Information Sciences 66 (2023) 202102.

[7] Z. Wu, C. Sun, H. Xuan, Y. Yan, Deep stereo video inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5693–5702.

[8] Z. Wu, C. Sun, H. Xuan, G. Liu, Y. Yan, Waveformer: Wavelet transformer for noise-robust video inpainting, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 6180–6188.

[9] Y. Zhang, Z. Wu, Y. Yan, Proceedings of the pfta-net: Progressive feature alignment and temporal attention fusion networks for video inpainting, in: IEEE International Conference on Image Processing, 2023, pp. 191–195.

[10] H. Xuan, Z. Wu, J. Yang, B. Jiang, L. Luo, X. Alameda-Pineda, Y. Yan, Robust audio-visual contrastive learning for proposal-based self-supervised sound source localization in videos, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2024) 4896–4907.

[11] H. Zhou, Q. Liu, Y. Wang, Learning discriminative representations for skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10608–10617.

[12] M. Contributors, Openmmlab's next generation video understanding toolbox and benchmark, https://github.com/open-mmlab/mmaction2, 2020.

[13] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7083–7093.

[14] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

[15] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[16] H. Duan, J. Wang, K. Chen, D. Lin, Pyskl: Towards good practices for skeleton action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7351–7354.

[17] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Transactions on Image Processing 29 (2020)

9532–9545.

[18] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.

[19] H. Duan, J. Wang, K. Chen, D. Lin, Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition, arXiv preprint arXiv:2210.05895 (2022).

[20] K. Li, D. Guo, G. Chen, X. Peng, M. Wang, Joint skeletal and semantic embedding loss for micro-gesture classification, arXiv preprint arXiv:2307.10624 (2023).

[21] H. Huang, X. Guo, W. Peng, Z. Xia, Micro-gesture classification based on ensemble hypergraph-convolution transformer, Micro-gesture Analysis for Hidden Emotion Understanding 2023 (2023) 9.

[22] F. Wang, D. Guo, K. Li, M. Wang, Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 5345–5353.

[23] F. Wang, D. Guo, K. Li, Z. Zhong, M. Wang, Frequency decoupling for motion magnification via multi-level isomorphic architecture, arXiv preprint arXiv:2403.07347 (2024).

[24] D. Guo, K. Li, B. Hu, Y. Zhang, M. Wang, Benchmarking micro-action recognition: Dataset, methods, and applications, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 6238–6252.

[25] K. Li, D. Guo, P. Liu, G. Chen, M. Wang, Mmad: Multi-label micro-action detection in videos, arXiv preprint arXiv:2407.05311 (2024).

[26] J. Zhou, D. Guo, Y. Zhong, M. Wang, Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling, arXiv preprint arXiv:2406.00919 (2024).

[27] Y. Wei, Z. Zhang, Y. Wang, M. Xu, Y. Yang, S. Yan, M. Wang, Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking, IEEE Transactions on Image Processing 30 (2021) 4788–4801.