Multi-modal Micro-gesture Classification via Multi-scale Heterogeneous Ensemble Network

Hexiang Huang[1], Yuhan Wang[1], Kerui Linghu[1] and Zhaoqiang Xia[1,2,*]

[1]*School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China*
[2]*Innovation Center NPU Chongqing, Northwestern Polytechnical University, Chongqing 400000, China*

Abstract

Micro-gesture classification has become an important research topic in the field of emotion analysis and human-computer interaction, and recently has received more and more attention. Although certain models of action recognition for normal behaviors have demonstrated promising results in classifying micro-gestures, these models still encounter significant challenges when processing micro-gestures that occur within subtle temporal windows. To end this, we propose a multi-scale heterogeneous ensemble network for micro-gesture classification with multi-modal data. This framework combines two models with different architectures and employs multi-scale residual connections within these models to capture fine-grained features and extend the range of receptive field. Simultaneously, we employ a novel data group training strategy, which can more effectively address the class-imbalance problem for model learning over the data. Finally, our model was evaluated on the iMiGUE dataset with Top-1 accuracy of **0.7019**, placing **second ranking** in the MiGA2024 Challenge (Track 1: Micro-gesture Classification).

Keywords

Micro-gesture classification, Multi-scale, Heterogeneous ensemble network, Class-imbalance

1. Introduction

Micro-gesture (MiG) classification refers to the process of recognition and classifying spontaneously occurring minute movements on the human face and body. The significance of MiG classification lies in its ability to capture and analyze the nuances of human behavior in detail, which has important value in multiple fields. In emotion analysis, MiG classification can provide more accurate detection of hidden emotional states, helping to better understand the user's real emotions. In human-computer interaction, MiG classification can improve the system's response accuracy and user experience, making the interaction process more natural and intelligent. In addition, in areas such as safety monitoring, health care and sports analytics, MiG classification can also provide critical insight and support. Therefore, MiG classification not only enriches the fine-grained research of behavior analysis, but also promotes the development and innovation of related applications.

Currently, the research about micro-gestures is relatively limited, which is different to the

MiGA2@IJCAI24: International IJCAI Workshop on 2nd Micro-gesture Analysis for Hidden Emotion Understanding, August 3, 2024, Jeju, Korea.

[*] Corresponding author.

huanghexiang@mail.nwpu.edu.cn (H. Huang); wwww@mail.nwpu.edu.cn (Y. Wang); linghu@mail.nwpu.edu.cn (K. Linghu); zxia@nwpu.edu.cn (Z. Xia)
0000-0003-0630-3339 (Z. Xia)

task of micro-expression recognition [1, 2] and the action recognition [3]. With the continuous development of action recognition algorithms [4], many models using different modalities have emerged. The MiG datasets such as SMG and iMiGUE [5, 6] encompasses a diverse array of multi-modal data types, which are characterized by their ability to integrate and represent different forms of information. However, in the initial phase of research conducted on the MiG data, the studies predominantly focused on the utilization of either RGB (Red, Green, Blue) imagery or skeleton modality data individually and often just transferring normal action classification algorithms directly to the task of MiG classification. To cite a few, the temporal segmentation network (TSN) [7], the temporal relation network (TRN) [8] and the temporal shift module (TSM) [4] have been proposed based on the RGB data, while spatio-temporal graph convolution network (ST-GCN) [9], multi-scale graph convolution (MS-G3D) [10] and enhanced hypergraph-convolution transformer (EHCT) [11, 12] have been presented with the skeleton data. While these methodologies are indeed valuable for conducting specific analyses, they may not fully exploit the extensive potential inherent in the multi-modal characteristics.

As the RGB data usually contains the color and texture information, it can capture the subtle changes of the human body under different lighting and background. However, this modality is greatly affected by environmental factors, which may lead to the deterioration of image quality and blur for fast-moving human bodies, affecting the recognition effect. Different from the RGB modality, skeleton data usually records the position coordinates of human joints, which can obtain the precise position of the human skeleton joint and is not affected by environmental factors. However, for the human body with partial occlusion or complex posture, the recognition accuracy may decline. Therefore, the proposed method can provide comprehensive information about human appearance and movements by combining the dual-modal data of RGB and skeleton [13]. Among them, the RGB modality can provide contextual information to help skeleton-modal data more accurately identify human posture and movements. The stability of skeleton-modal data can make up for the lack of RGB-modal data affected by environmental factors.

To more effectively capture micro-reactions within MiGs, we construct a Res2Net3D structure utilizing the multi-scale residual architecture from Res2Net [14]. The Res2Net framework enhances feature extraction across different scales by incorporating multi-scale residual modules, allowing the simultaneous processing of local and global information and thereby augmenting the model's feature representation capability. This multi-scale feature representation enables the model to detect subtle changes and details across various scales, which is crucial for identifying MiG categories. By extending the Res2Net architecture to a 3D version (Res2Net3D), the model can process spatiotemporal information and capture dynamic gesture patterns. This extension is vital for handling time-dependent information in MiG classification, facilitating the identification of continuous micro-gesture variations. In addition, we find that a single network may overfit a particular data distribution and not generalize well to other scenario (i.e., other subjects). By fusing different network structures, this problem can be alleviated and the generalization ability of the model can be improved, making its performance on different settings and tasks more stable. Since different network architectures have their own advantages and advantages, these advantages can be comprehensively utilized in one model to improve the overall performance by combining multiple structures. Therefore, we ensemble heterogeneous sub-networks on the basis of our previous model [11] for MiGA2023 challenge, and continue to achieve significant performance improvement on the iMiGUE dataset [6]. The main contributions of this paper can be summarized as:

- We propose a deep framework with multi-modal and multi-scale heterogeneous ensemble network (M2HEN) for the task of MiG classification, capturing the diversity of data and enhancing the representation of the model..
- We design a multi-scale residual module in 3D structure to improve the sensitivity of the model for micro-gestures.
- We employ a novel data group training strategy, which can more effectively address the class-imbalanced problem in the data.
- We perform extensive experiments and achieve the second ranking in the Track 1 of MiGA2024 Challenge.

2. Methodology

The main framework of our proposed method (M2HEN) is shown in Fig. 1. In the framework, we construct a heterogeneous ensemble network, using two deep models with completely different structures (one is based on 3D convolution, and the other is based on Transformer) for ensemble learning. By designing this heterogeneous ensemble model, we can increase the diversity of features and improve the representation ability of the model. For the base model, we propose MiG-enhanced Multi-modal and Multi-scale 3D Convolutional sub-Network (M3CN) as the 3D convolution model and Ensemble Hypergraph-Convolution Transformer (EHCT) [11] as the Transformer model.

2.1. Multi-modal and Multi-scale 3D Convolutional Network

The M3CN sub-network uses both skeleton-modal and RGB-modal data. Inspired by PoseC3D [13], we transform the raw skeleton data $\vec{S^T} = \left\{\vec{S_1^T}, \vec{S_2^T}, ..., \vec{S_t^T}\right\}$ into a 3D heatmap volume with the size of $n \times t \times H \times W$, where $\vec{S_i^T} = \{\vec{s_1}, \vec{s_2}, ..., \vec{s_n}\}$, n is the number of key points, t is the number of frames in a clip, H and W are the height and width of the heatmap. Through the coordinates of skeleton joints $\vec{s_j} = (x, y, c)$, joint heatmap J can be obtained by combining n Gaussian mappings centered on each joint:

$$J_{nij} = e^{-\frac{(i-x)^2+(j-y)^2}{2*\sigma^2}} * c, \tag{1}$$

where the parameter σ represents the variance of the Gaussian graph, and (x, y) and c are the position coordinates and confidence score of the n-th joint, respectively.

To be able to better capture the subtle and detailed gesture changes in one video, we extend the Res2Net [14] into 3D version (Res2Net3D) as the backbone network for feature extraction, which induces the multi-scale information compared to the ResNet3D. In the bottleneck module of Res2Net3D, after the first 3D convolution with a kernel of $1 \times 1 \times 1$, the obtained feature map $X \in \mathbb{R}^{C,T,H,W}$ is evenly divided into 4 feature maps $x_1, x_2, x_3, x_4 \in \mathbb{R}^{C/4,T,H,W}$ in the channel dimension, and each feature map is processed by convolution and residual fusion, so as

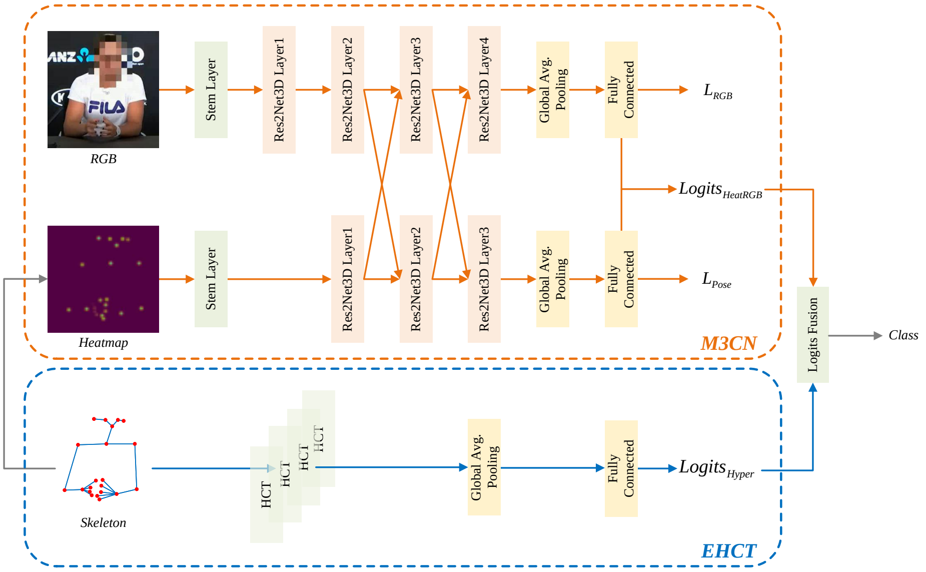

Figure 1: The overall framework of the proposed method: the upper portion comprises of the MiG-enhanced Multi-modal and Multi-scale 3D Convolutional sub-Network (M3CN), which utilizes RGB and heatmap data as input, while the lower portion consists of the Ensemble Hypergraph-Convolution Transformer (EHCT) model, which mainly leverages the skeleton data.

to obtain fusion features with different scale information $y \in \mathbb{R}^{C,T,H,W}$. The specific formula is shown in Eq. 2:

$$\begin{aligned} y = concat(x_1, Conv(x_2), Conv(x_3 + Conv(x_2)),\\ Conv(x_4 + Conv(x_3 + Conv(x_2))), axis = 1), \end{aligned} \tag{2}$$

where $Conv$ denotes the 3D convolution with a kernel of $1 \times 3 \times 3$ and $concat\ (\ldots, axis = 1)$ denotes the concatenation of the inner elements in the second dimension (the channel dimension).

In terms of multi-modal data feature fusion, M3CN uses a two-branch structure similar to SlowFast[15], the RGB branch uses a smaller frame number and a larger channel number, and the Skeleton branch uses a larger frame number and a smaller channel number. As shown in Fig. 1, the outputs of Res2Net3D Layer2 and 3 in the RGB branch are cross-fused with the outputs of Res2Net3D Layer1 and 2 in the Skeleton branch, and are equally fused in the final fully connected layer of both.

2.2. Ensemble Hypergraph-Convolution Transformer

The Ensemble Hypergraph-Convolution Transformer (EHCT) model only uses raw skeleton data as input. The input $\vec{S_i^T} = \{\vec{s_1}, \vec{s_2}, ..., \vec{s_n}\}$ represents n key points extracted from frame i, including those pertaining to the body, face, left and right hands, are presented in 2D format $\vec{s_j} = (x, y, c)$ by using the protocol of OpenPose [16]. In the EHCT model, the hypergraph-convolutional Transformer and the enhanced hypergraph self-attention mechanism are used to process these key points, and then the main classifier and auxiliary classifier are used for multi-branch ensemble learning. Finally, the output of the two classifiers is combined to obtain the classification results of recognizing micro-gestures.

In the self-attention module of EHCT shown in Fig. 1, the feature E_f with the hyperedges of hypergraph is constructed by Eq. 3:

$$E_f = HD_e^{-1}H^TSW_e, \tag{3}$$

where H represents the incidence matrix of key points and hyperedges. In the matrix H, each row represents a key point and each column represents a hyperedge. D_e is the diagonal matrix representing the degree matrix of hyperedges, and W_e represents the projection matrix of hyperedges. For more details on EHCT, please see the previous work [11].

2.3. Training Strategy

The MiG data used in our study usually exhibits a long-tail distribution, as illustrated in Fig. 2 (Taking iMiGUE for example). In our prior work of EHCT [11], we has proposed the utilization of primary and auxiliary classifiers to mitigate imbalanced classes. Building upon this, we partition the imbalanced dataset into four parts ($Part\ A$, $Part\ B$, $Part\ C$, and $Part\ D$) and trained the model using various combinations of these subsets on each occasion. This approach effectively enhances the accuracy of tail classes and consequently elevates the overall classification performance.

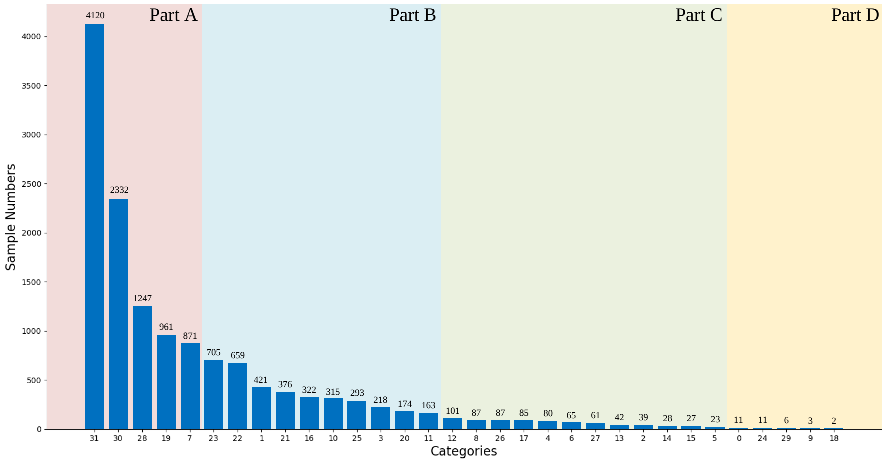

Figure 2: Sample distribution of iMiGUE dataset and 4 parts divided by the limit of 1/5 of the maximum sample numbers of the current part.

The allocation of the dataset for training the model is as follows:

$$Trainset_i = \sum \underbrace{Part\{A, B, C, D\}}_{i} + Other_i, 1 \leq i \leq 4, \tag{4}$$

where $Trainset_i$ means that the model is trained using i parts of the data (eg. $Trainset_3 = Part\ B + Part\ C + Part\ D + Other_3$) and $Other_i$ means an even selection of a certain number of categories from the currently unused part (head categories) as a category called "OTHERS", which can effectively prevent the accuracy of the head categories from decreasing. In addition, the training set when $i = 4$ is called the primary training set, and the training set when $i < 4$ is called the tail training set.

Since dividing the training set into different parts changes the label distribution, when the model predicts a non-OTHERS category, the labels of the categories in the tail training set are mapped one-to-one to the original labels in the primary training set. With the logits from different classifiers which use different training data, the way of combining these outputs is calculated as follows:

$$Logits = \sum_{i=1}^{4} \alpha_i \cdot Map\{Logits_{HeatRGB_i} + Logits_{Hyper_i}\}, \tag{5}$$

where the hyperparameter α_i denoted as the weight by which the logits of the i-th classifier, and Map means when the model predicts a tail category, it is weighted into the primary logit by using a mapping relationship between labels.

3. Experiments

In this section, we evaluate our model on the iMiGUE dataset [6] by following the protocol of MiGA2024 Challenge (Track 1: Micro-gesture Classification). The dataset, metrics, ablation study and comparison experiments are reported in the following subsections.

3.1. Dataset and Metrics

In this challenge, the iMiGUE [6] dataset with fixed training and test samples is used to evaluate our proposed method. This dataset includes a total of 32 categories of MiGs, and covers two emotions as well as 72 subjects with each gender accounting for half of the total number of subjects. It consists of 18,499 samples taken from 359 videos with a resolution of 1280×720. Each video is about 0.5-25.8 minutes long. Since the iMiGUE dataset is collected in-the-wild setting, the overall dataset presents a long-tailed (unbalanced) distribution.

To evaluate the classification performance of our model, we employ Top-1, Top-5 and Class Average Accuracy as evaluation metrics, the equations of the metrics are as follows:

$$Acc_{Top-1} = \frac{\sum_{i=1}^{N}[argmax(P(y_i|x_i)) = y_i]}{N}, \tag{6}$$

$$Acc_{Top-5} = \frac{\sum_{i=1}^{N}[y_i \in top5(P(y_i|x_i))]}{N}, \tag{7}$$

$$Acc_{cls\_avg} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} Acc_{Top-1_i}, \tag{8}$$

where N denotes the number of samples, $\mathcal{N}$ denotes the number of categories, x_i denotes the feature of the i-th sample, y_i denotes the true label of the i-th sample, $P(y_i|x_i)$ denotes the probability distribution obtained from the model's predictions for the i-th sample, $top5$ denotes the top five categories with the highest probabilities, and cls_avg denotes the average accuracy for each category.

3.2. Ablation Study

Table 1

Performance comparison on iMiGUE dataset, where NoC, HgE and GpT denote the number of clips, heterogeneous ensemble and group training, respectively.

Backbone	Modality	NoC	HgE	GpT	Accuracy(%)		
					Top-1	Top-5	Class Avg.
ResNet3D	RGB	1	✘	✘	57.62	87.07	27.49
	Skeleton	1	✘	✘	61.73	89.19	34.64
	RGB+Skeleton	1	✘	✘	66.59	92.79	41.09
		5	✘	✘	67.27	93.07	41.59
		10	✘	✘	67.27	93.07	41.35
		15	✘	✘	67.16	93.16	41.13
Res2Net3D	RGB	1	✘	✘	61.14	89.35	28.07
	Skeleton	1	✘	✘	61.11	87.72	35.95
	RGB+Skeleton	1	✘	✘	66.57	91.98	41.63
		5	✘	✘	67.69	92.53	42.12
		10	✘	✘	68.11	92.63	41.44
		15	✘	✘	68.17	92.66	41.66
		10	✔	✘	69.47	93.64	40.29
		10	✔	✔	**70.19**	**93.69**	**47.36**

Firstly, in order to verify the effectiveness of the proposed model, we conduct a series of ablative experiments, and the specific results can be obtained from Table 1. We employ ResNet3D as the baseline network and train it using both single-modal and multi-modal data. It is observed that the dual-modality fusion approach indeed significantly enhances the performance of the model. Based on this, in order to more accurately capture the subtle movements and changes in MiGs, we adopt Res2Net3D as the backbone network. The experimental results show that this multi-scale residual structure significantly improves the sensitivity and accuracy of the model for MiG classification. Numerically, this improved method promotes the accuracy of the model by 0.84%.

Since the frame length of the input data is different, the model needs to be down-sampled to the required number of frames, which will produce a certain loss of information. Therefore, different NoC are experimented, in which the selected frame index for each clip is different from each other, which can effectively alleviate this problem. In order to capture the diversity of data and enhance the representation capability of the model, we further construct a heterogeneous ensemble model using 3D CNN and Transformer. Compared with the single model, the accuracy of the proposed method increases by 1.3% for Top-1 and 0.98% for Top-5.

In terms of dealing with class-imbalanced problem, we use the strategy of group training to effectively alleviate the long-tail effect of the model, and the Top-1 accuracy of the model is increased by 0.72%, and the average accuracy of the class is increased by 7.07%.

3.3. Comparison to State-of-the-art Methods

Table 2
The comparison results of various methods on iMiGUE dataset.

Methods	Model+Modality	Accuracy(%)	
		Top-1	Top-5
ST-GCN [9]	GCN + Skeleton	46.97	84.09
MS-G3D [10]		54.91	89.98
TSN [7]	2DCNN + RGB	51.54	85.42
TRN [8]		55.24	89.17
TSM [4]		61.10	91.24
Hyperformer [17]	Transformer + Skeleton	57.01	87.86
EHCT [11]		63.02	91.36
PoseC3D [13]	3DCNN + RGB	57.62	87.07
	3DCNN + Skeleton	61.73	89.19
	3DCNN + (RGB + Skeleton)	67.27	93.16
M2HEN(Ours)	(3DCNN + Transformer) + (RGB + Skeleton)	**70.19**	**93.69**

Our proposed method is also examined through a comparative analysis on iMiGUE dataset, which is shown in Table 2. We compare our proposed method with state-of-the-art methods. Compared with the single-modal EHCT [11], we introduce a multi-modal model of RGB image and skeleton data, which significantly improves the accuracy by 7.17%. Furthermore, compared to the RGBPoseC3D[13] model, which also uses multi-modal inputs, we used a heterogeneous ensemble network, and this innovative architecture design enabled us to improve the accuracy of our model again by 2.92%.

4. Conclusions

In conclusion, in the study of micro-gesture (MiGs) classification, we significantly improved the accuracy of MiG classification through a series of innovative techniques. We designed a 3D multi-scale residual module to improve the sensitivity of the model to small changes in MiGs. A heterogeneous ensemble network was constructed to enhance the ability of data diversity capture and model characterization. A novel data grouping training strategy was implemented to effectively solve the class-imbalanced problem. The comprehensive application of these strategies not only optimized the performance of the model, but also layed a foundation for future research on MiG classification.

References

[1] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, IEEE Transactions on

Image Processing (2020) 8590–8605.

[2] X. Guo, X. Zhang, L. Li, Z. Xia, Micro-expression spotting with multi-scale local transformer in long videos, Pattern Recognit. Lett. (2023) 146–152.

[3] W. Peng, J. Shi, Z. Xia, G. Zhao, Mix dimension in poincaré geometry for 3d skeleton-based action recognition, ACM International Conference on Multimedia (ACM MM) (2020) 1432–1440.

[4] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, International Conference on Computer Vision (ICCV) (2019) 7082–7092.

[5] H. Chen, X. Liu, X. Li, H. Shi, G. Zhao, Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning, IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (2019) 1–8.

[6] 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, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 10626–10637.

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. V. Gool, Temporal segment networks for action recognition in videos, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 2740–2755.

[8] B. Zhou, A. Andonian, A. Torralba, Temporal relational reasoning in videos, European Conference on Computer Vision (ECCV) (2018) 831–846.

[9] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, AAAI Conference on Artificial Intelligence (AAAI) (2018) 7444–7452.

[10] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 140–149.

[11] H. Huang, X. Guo, W. Peng, Z. Xia, Micro-gesture classification based on ensemble hypergraph-convolution transformer, IJCAI Workshop&Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) (2023) 1–9.

[12] X. Guo, W. Peng, H. Huang, Z. Xia, Micro-gesture online recognition with graph-convolution and multiscale transformers for long sequence, IJCAI Workshop&Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) (2023) 1–8.

[13] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition., IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2959–2968.

[14] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 652–662.

[15] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, International Conference on Computer Vision (ICCV) (2019) 6201–6210.

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) 172–186.

[17] Y. Zhou, C. Li, Z.-Q. Cheng, Y. Geng, X. Xie, M. Keuper, Hypergraph transformer for skeleton-based action recognition, arXiv abs/2211.09590 (2022).