Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/54/20241692

Work procedure action recognition based on skeleton and video features

Yu-Chen Cai[1,2], Zhi-Cong Zhang[1]

[1]School of Mechanical Engineering, DongGuan University of Technology, Donguan 523808, China

[2]caiyuchen999@163.com

Abstract. In the field of industrial engineering, traditional methods for analyzing manufacturing process actions have limitations such as time-consuming, labor-intensive, and experience-dependent. To address these challenges in action analysis, we propose an intelligent action recognition method based on both skeleton and video features, aiming to replace manual decomposition of action elements. The MediaPipe framework is used for human posture estimation to obtain the skeleton sequence, and the CNN-GRU model is constructed action recognition based on skeleton features. For hand movements involving the use of industrial gloves, an enhanced TimeSformer video understanding model is introduced for action recognition based on video features. This improvement incorporates uniform attention and external attention mechanisms, resulting in enhanced model performance. The final experimental validation for the self-constructed process action dataset shows that the online detection speed of the skeleton action recognition model reaches 25 FPS, and the accuracy of the end-to-end video action recognition model is improved by 10.5 percentage points compared to the base model.

Keywords: Industrial Engineering, Action Recognition, Skeleton Features, Video Features, Attention

1. Introduction

In manufacturing systems, analyzing frontline operators' process actions is crucial for improving industrial engineering production processes. Stemming from Gilbreth's pioneering work in motion study, action analysis involves observing and recording actions, followed by decomposing action elements based on operator movement sequences. Traditional methods rely on manual on-site observation and action decomposition, which is time-consuming, labor-intensive, and prone to high error rates due to subjectivity. Research on intelligent action recognition is crucial for improving management and production efficiency in manufacturing. Common production process actions include: carry, dismantle, delay, check, preset, operate, move, and others. In the deep learning era, significant progress has been made in action recognition models, with most exhibiting excellent performance on public datasets.

Human pose estimation technology involves detecting key points in the human skeleton. Cao et al. [1] introduced the Openpose model, which employs a bottom-up approach to cluster key points. Google Research released a human pose estimation method based on MediaPipe, and Blaze Pose, proposed by Bazarevsky et al. [2], relies on a single camera to infer the coordinates of 33 3D body keypoints. Zhang

DOI: 10.54254/2755-2721/54/20241692

et al. [3] proposed the Mediapipe Hands model, which infers 21 3D keypoints for hand tracking. Behavior recognition is a fundamental task in the field of video understanding. Carreira et al. [4] introduced the I3D model, a two-stream model based on 3D convolutions for video feature extraction. 3D networks have dominated the video understanding field since then, until the emergence of the Vision Transformer (ViT) [5]. TimeSformer[6] was the first model to apply the Vision Transformer to video understanding, introducing a separation of spatiotemporal self-attention. Google Research also made improvements to the ViT model, proposing the ViViT model[7], a Transformer-based video classification model.

2. Framework of action recognition tasks

In the industrial domain, there is a lack of relevant action datasets specifically tailored to production and manufacturing scenarios. Additionally, there is a scarcity of experimental research on the practical application of models in the industrial sector. Furthermore, existing datasets for most human pose estimation models do not include images of hands wearing industrial gloves. Consequently, when performing process action recognition based on skeleton sequences, hand actions involving glove-wearing cannot be accurately identified through hand key points. To address these issues, this paper proposes a sequential detection approach for process action based on both skeleton sequences and video features. The approach involves collecting data for different process actions based on commonly used ergonomic action elements, and a schematic framework is illustrated in figure 1.

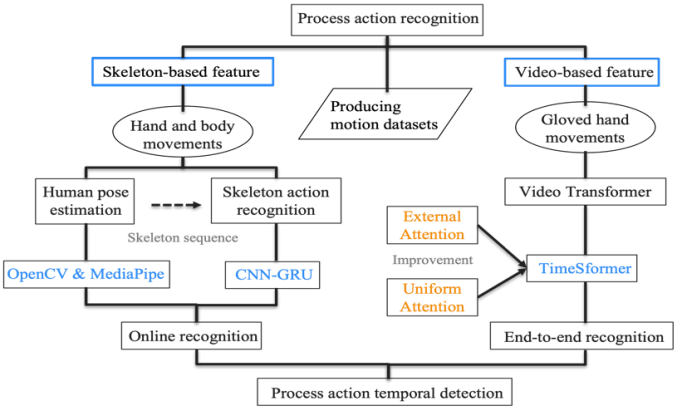

Figure 1. Process action recognition task framework.

The main contributions of this paper are summarized below: (1) Creating a dataset for process action in the industrial domain: The dataset is generated by collecting skeleton data and video data for different process actions based on commonly used ergonomic action elements in human factors engineering. The research focuses on 12 classes of dynamic feature process actions, including 8 effective actions, 2 ineffective actions, and 2 auxiliary actions. (2) Constructing a skeleton action recognition model: The Holistic solution in MediaPipe is employed for human pose estimation to obtain skeleton sequences. Ultimately, a CNN-GRU network is utilized for the classification of process actions. (3) Constructing a video action recognition model: To address glove-wearing hand action recognition challenges, an end-to-end video classification model, TimeSformer, is created. External and uniform attention mechanisms are introduced for enhanced performance.

3. Construction Methods of Work Procedure Action Recognition Model

Utilizing a standard 2D camera in conjunction with the MediaPipe framework for human pose estimation, the acquired three-dimensional skeleton keypoint coordinates are fed into a CNN-GRU network for action classification. For video data, pre-trimmed video segments are input into the enhanced TimeSformer model for training. Ultimately, the model predicts action categories frame by frame within the video stream.

DOI: 10.54254/2755-2721/54/20241692

3.1. Skeleton-based action recognition

3.1.1. Human pose estimation. Developing a human pose estimation model involves three solutions, and the workflow is illustrated in figure 2. MediaPipe Pose initially utilizes a face-based human detection system to locate regions of interest (ROI) within each frame. These ROIs are then cropped and used as input for the human keypoint detector, predicting the coordinates of body keypoints within the ROIs. MediaPipe Hands focuses on hand keypoints, obtaining their coordinates through a palm detection and hand keypoints detection system. MediaPipe Holistic integrates body and hand keypoints, creating a semantically consistent end-to-end solution while simultaneously inferring multiple neural networks. The process involves using MediaPipe Pose for keypoint detection, obtaining 33 body keypoints, exporting 3 ROIs for each hand, and enhancing the ROIs using a hand-specific cropping model. Subsequently, MediaPipe Hands infers 42 hand keypoints. Finally, the keypoints from both the hand and body models are fused to derive a total of 75 keypoint coordinates.

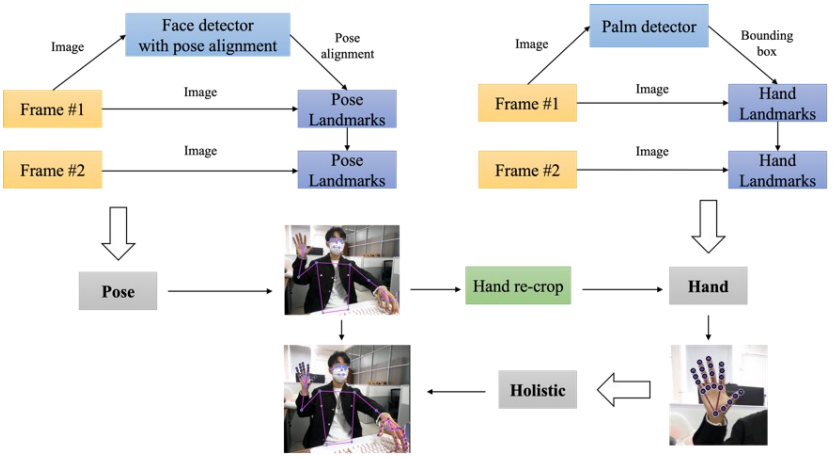

Figure 2. Workflow for human posture estimation.

3.1.2. CNN-GRU. To enhance real-time performance and reduce computational resource utilization, a CNN-GRU action classidfication model based on skeleton sequences is proposed. The model predicts action categories frame by frame in real-time video streams, constructing action features by considering both historical and current temporal information. In this model, the GRU architecture combines the input gate and forget gate of the LSTM model into an update gate, controlling the amount of past memory information retained for the current moment's data. The CNN-GRU model built in this study is a hybrid predictive model based on convolutional neural networks and gated recurrent units, as depicted in table 1. Initially, features are extracted using CNN, followed by utilizing the GRU neural network to learn the dynamic patterns of feature variations for prediction. On one hand, leveraging the powerful feature extraction capability of CNN, intrinsic connections between skeleton data are explored, thereby reducing the scale and complexity of the original data. On the other hand, the temporal memory capacity of the GRU neural network is employed to learn the dynamic patterns within the skeleton data, establishing a nonlinear relationship between input and output.

Table 1. Model structure of CNN-GRU

Layer	Output shape	Param	Connected to
Input Layer	(None, 30, 258)	0	
Conv1D	(None, 28, 64)	49600	Input Layer
MaxPooling1D	(None, 14, 64)	0	Conv1D
Flatten	(None, 896)	0	MaxPooling1D
Dense	(None, 64)	57408	Flatten

DOI: 10.54254/2755-2721/54/20241692

Table 1. (continued).

Reshape	(None, 64, 1)	0	Dense
GRU	(None, 64, 100)	30900	Reshape
GRU_1	(None, 100)	60600	GRU
Dense_1	(None, 64)	6464	GRU_1
Concatenate	(None, 128)	0	Dense Dense_1

3.2. Video-based action recognition

3.2.1. TimeSformer. In order to recognize hand movements during the process of wearing work gloves, improvement and training were conducted based on the TimeSformer model. As depicted in figure 3, the model architecture of TimeSformer is illustrated. The backbone network utilizes the Vision Transformer network, and a separate spatiotemporal attention mechanism is constructed to reduce computational load. As shown in figure 3a, the video segment is initially inputted, and each frame is divided into multiple image patches. These two-dimensional image patches are transformed into one-dimensional embedding vectors through a linear mapping layer. After concatenating the classification tokens and position encoding, the result is passed to the Transformer encoder. In figure 3b, the tokens undergo L layers of encoder, with each encoder layer involving multi-head self-attention(MSA), layer normalization(LN), and a multi-layer perceptron(MLP). Figure 3c illustrates the separate spatiotemporal attention mechanism, where temporal self-attention is first applied to all image patches within the same frame, followed by spatial self-attention to the corresponding positions of image patches in different frames. Finally, the classification head MLP Head processes the class token for classification symbols, yielding the ultimate prediction results.

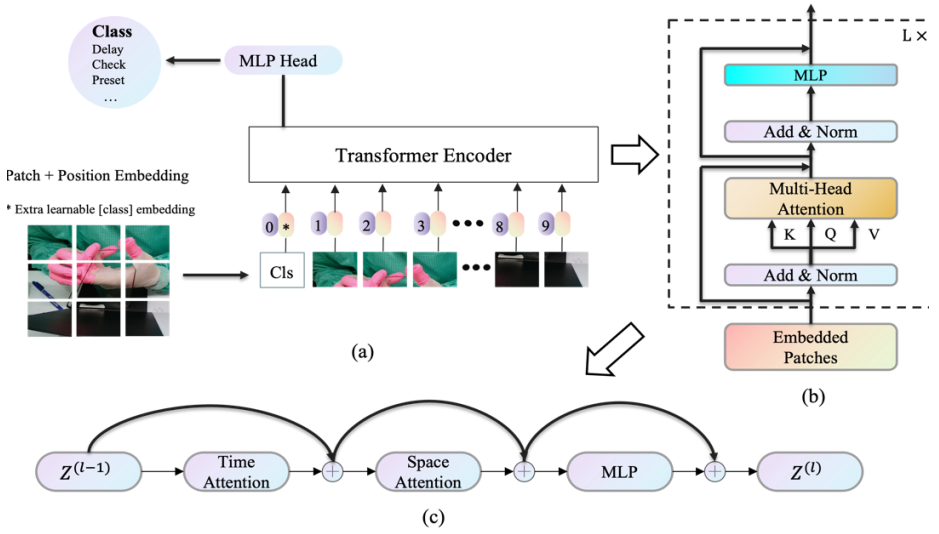

Figure 3. Model architecture of TimeSformer. (a) illustrates the architecture of the Vision Transformer model. (b) showcases the internal structure of the Transformer encoder. (c) depicts the separate spatiotemporal attention.

3.2.2. External Attention and Uniform Attention. In the TimeSformer model, two new attention mechanisms have been introduced, primarily enhancing the relevant modules in the Transformer encoder. As depicted in figure 4b, External attention[8] serves as an improvement to self-attention in the original Transformer structure[9], making the Transformer, initially designed for the NLP domain, more applicable to computer vision tasks. It reveals the intrinsic connection between linear layers and

DOI: 10.54254/2755-2721/54/20241692

attention mechanisms, demonstrating that linear transformations are, in fact, a specific form of attention implementation, leading to reduced computational complexity. Furthermore, based on two shared external memory units, the model implicitly learns features of the entire dataset rather than individual sample features, thereby enhancing the understanding of video context.

Due to the self-attention mechanism employed by the ViT model to establish internal relationships within images, the model exhibits a strong preference for dense interactions. Uniform attention[10] assists the model in better understanding and processing different regions within an image. Figure 4c introduces the Context Broadcasting module (CB) at the MLP's end to manually incorporate uniform attention, thereby reducing the attention map density. Without the CB module, the ViT model must autonomously learn effective attention mechanisms, possibly resulting in insufficient attention to specific regions. Introducing the CB module guides attention allocation in each layer, promoting a more comprehensive focus on various image parts and facilitating improved video information learning.

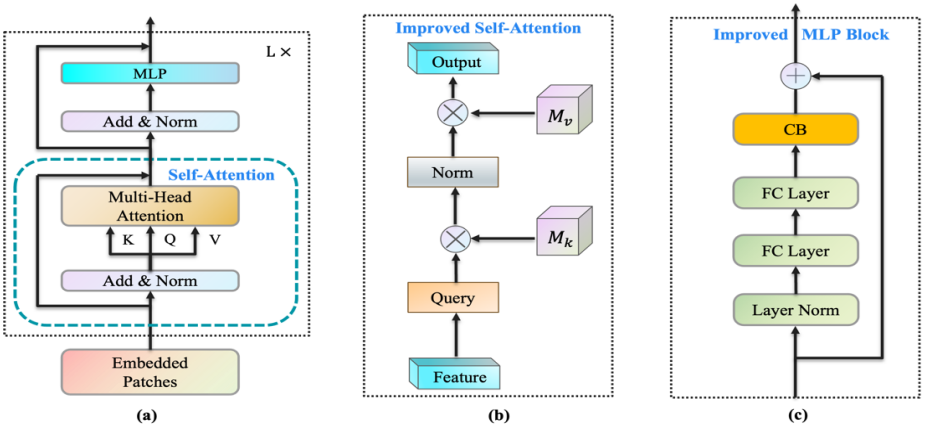

Figure 4. External attention and Uniform attention. (a) represents the Transformer encoder. (b) illustrates External Attention. (c) depicts Uniform Attention.

4. Dataset Creation and Model Training

4.1. Creation of the Action Dataset for Work Procedures

The experimental environment setup for data collection is as follows: Intel Core i7-10750H 2.60GHz, 6 cores 12 threads, 16GB memory, Windows 10 64-bit system, and a monocular RGB camera. The programming language used is Python 3.8, and the frameworks include OpenCV 4.6.0 and MediaPipe 0.8.9. For skeleton action recognition, the experiment involved collecting skeleton sequences of different work procedures in various scenarios to create the dataset. Ten different work procedures were simulated and encoded according to the sequence of action categories: carry, move, select, preset, delay, stretch, check, operate, locate, and hold. Each action was recorded in 40 videos, with 30 consecutive frames per video used to generate dynamic action sequences. Each file includes 258 keypoint data, resulting in 7740 keypoint data for each action. In total, 400 videos and 12,000 data points were generated. The training set consists of 8400 data points, while the test set contains 3600 data points.

For video action recognition, videos capturing hand actions of wearing gloves were recorded for an hour. The videos were then segmented and categorized into six different hand actions of wearing gloves: dismental, delay, check, use, rotation, and preset. The number of samples for each action is as follows: 123 for dismental, 70 for delay, 251 for check, 78 for use, 233 for rotation, and 123 for preset. The videos for each category were shuffled and divided into training and validation sets, with a ratio of 0.8 for the training set and 0.2 for the validation set.

DOI: 10.54254/2755-2721/54/20241692

4.2. Training the Skeleton Action Recognition Model

The experimental environment configuration for training the skeleton action recognition model is as follows: NVIDIA GeForce RTX 3060 with 8GB VRAM, TensorFlow 2.8.0. The CNN-GRU network's first layer has an input dimension of (30, 258), indicating that each video consists of 30 frames, and each frame has 258 keypoints. The final model concatenates the fully connected layers from both the CNN and GRU networks, adjusting the output dimension as needed, and utilizes the softmax function for action classification. The loss function employed is the cross-entropy loss, and the Adam optimizer is chosen for continuously updating network weights. The experiments were conducted using a self-constructed dataset simulating eight different work procedures. The evaluation metric utilized is the multi-class accuracy function, and the confusion matrix is employed to assess the accuracy of action classification for different work procedures. As shown in figure 5, a comparison among three different temporal models was performed. The horizontal axis represents predicted labels, while the vertical axis represents true labels. The LSTM network achieved a 97.7% accuracy after 180 iterations, with a total parameter count of 237,416. The GRU network reached a 99.2% accuracy after 80 iterations, with a total parameter count of 180,456. The CNN-GRU network achieved 100% accuracy after 80 iterations, with a total parameter count of 209,364. Therefore, the CNN-GRU model demonstrated the best performance in the work procedure action dataset based on skeleton features.

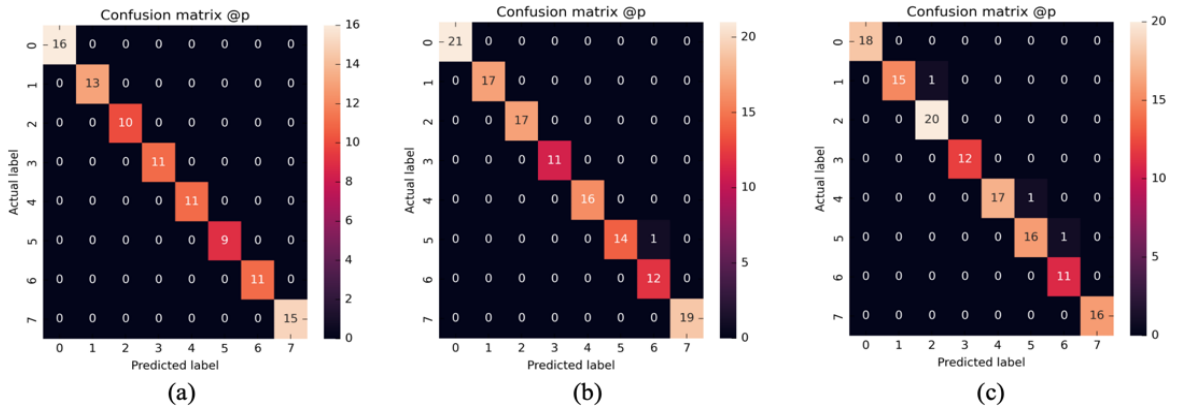

Figure 5. Confusion matrix of skeleton action recognition models. (a) depicts the confusion matrix for the CNN-GRU model. (b) illustrates the confusion matrix for the GRU model. (c) represents the confusion matrix for the LSTM model.

4.3. Training the Video Action Recognition Model

The experimental environment configuration for training the video action recognition model is as follows: Ubuntu 16.04 operating system, Intel(R) Xeon(R) Silver 4214R 2.40GHz processor, 256GB DDR4 1600MHz RAM, and 4 NVIDIA GeForce RTX 3090 GPUs with 24GB VRAM each. Key environment configurations include CUDA11.4, cuDNN8.2.4, and PyTorch1.13.0. During the experiments, adjustments were made to relevant parameters, setting the learning rate to 0.005, weight decay to 0.0001, and batch size to 8. Batch data loading was performed through 4 working processes. Each video was sampled with 16 frames, and the time interval between sampled frames was set to 16 frames. In the comparative experiments, additional testing was conducted on the ViViT model using a self-constructed dataset of video work procedure actions. As depicted in figure 6a, the accuracy of the ViViT model is slightly lower compared to TimeSformer. However, with the incorporation of external attention and uniform attention, the accuracy of the ViViT model improved by 2.5 percentage points. As shown in figure 6b, the TimeSformer model achieved the highest accuracy after integrating the improved attention mechanisms, resulting in a 10.5 percentage point improvement over the base model. The proposed external attention mechanism incorporates two memory units using shared parameters across the entire dataset, implemented through linear layers. Strong regularization enhances attention generalization. The CB module introduces essential uniform attention to each layer of the ViT model,

DOI: 10.54254/2755-2721/54/20241692

addressing the challenging yet crucial dense interactions required by the ViT model. In the manufacturing context, where extensive datasets on work procedure actions are scarce, this study focuses on the noteworthy performance enhancement achieved by integrating improved attention mechanisms into the model.

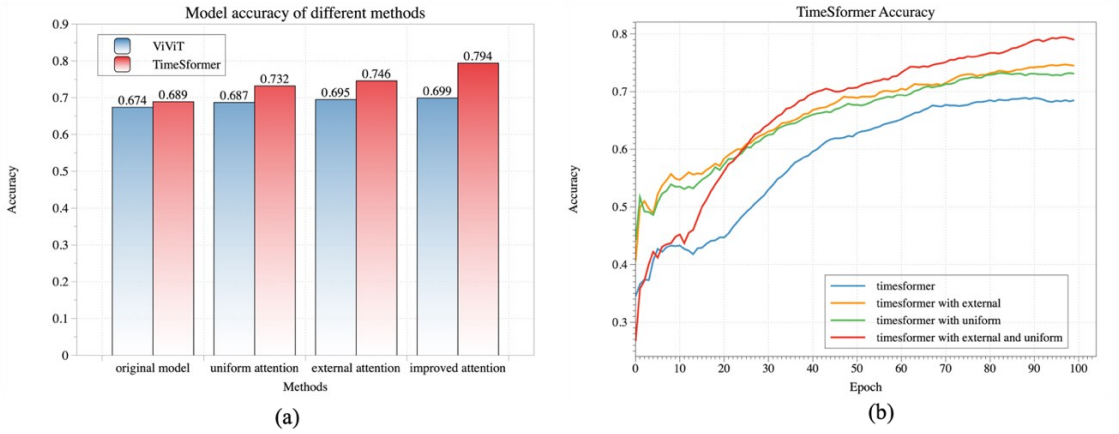

Figure 6. Accuracy of video action recognition models. (a) compares accuracy between TimeSformer and ViViT. (b) depicts the accuracy trend during TimeSformer training.

5. Conclusion

This study addresses limitations in traditional industrial engineering for work procedure analysis, renowned for its time and labor intensiveness. We focus on intelligent recognition using skeleton and video features, creating independent datasets for each format. A human pose model generates skeleton sequences, and a CNN-GRU network classifies dynamic features. Despite high prediction accuracy and online recognition speed (25FPS on CPU), this approach struggles to detect key points on gloved hands. To address this, an end-to-end TimeSformer for recognizing gloved hand actions is developed, enhancing performance with external and uniform attention. The ViViT model sees a 2.5 percentage point accuracy improvement with additional attention. The TimeSformer, with two attention mechanisms, exhibits a 10.5 percentage point accuracy increase over the base model, showcasing optimal performance on the work procedure action dataset. This study achieves intelligent recognition of work procedures, offering innovative solutions for manufacturing process optimization.

References

[1] Cao Z, Hidalgo G, Simon T, Wei S E and Sheikh Y 2021 IEEE Trans Pattern Anal Mach Intell. 43 172-186
[2] Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F and Grundmann M 2020 Blazepose: On-device real-time body pose tracking Preprint cs/2006.10204
[3] Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang C L and Grundmann M 2020 Mediapipe hands: On-device real-time hand tracking Preprint cs/2006.10214
[4] Carreira J and Zisserman A 2017 IEEE Conf. on Computer Vision and Pattern Recognition (Hawaii: Electrical and Electronics Engineers) pp 4724-33
[5] Dosovitskiy A, et al. 2020 An image is worth 16x16 words: Transformers for image recognition at scale Preprint cs/2010.11929
[6] Bertasius G, Wang H and Torresani L 2021 Proc. of the 38th Int. Conf. on Machine Learning vol 139(Vienna: Proc. of Machine Learning Research) pp 813-824
[7] Anurag A, Dehghani, Heigold G, Sun C, Lucic M and Schmid C 2021 IEEE/CVF Int. Conf. on Computer Vision(Canada: Electrical and Electronics Engineers) pp 6816-26
[8] Guo M H, Liu Z N, Mu T J and Hu S M 2022 *IEEE Trans Pattern Anal Mach Intell*. **45** 5436-47

DOI: 10.54254/2755-2721/54/20241692

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, N.Gomez A, Kaiser L and Polosukhin I 2017 *Adv Neural Inf Process Syst*. **30**
[10] Hyeon-Woo N, Yu-Ji K, Heo B, Han D, Oh S J and Oh T H 2023 Proc. of the IEEE/CVF Int. Conf. on Computer Vision(Paris) pp 5807-18