2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Occluded Part-aware Graph Convolutional Networks for Skeleton-based Action Recognition

Min Hyuk Kim, Min Ju Kim, Seok Bong Yoo*

Abstract— **Recognizing human action is one of the most critical factors in the visual perception of robots. Specifically, skeleton-based action recognition has been actively researched to enhance recognition performance at a lower cost. However, action recognition in occlusion situations, where body parts are not visible, is still challenging. We propose an occluded part-aware graph convolutional network (OP-GCN) to address this challenge using the optimal occluded body parts. The proposed model uses an occluded part detector to identify occluded body parts within a human skeleton. It is based on an autoencoder trained on a nonoccluded human skeleton and exploits the symmetry and angular information of the skeleton. Then, we select an optimal group constructed considering the occluded body parts. Each group comprises five sets of joint nodes, focusing on the body parts, excluding the occluded ones. Finally, to enhance interaction within the selected groups, we apply an interpart association module, considering the fusion of global and local elements. The experimental results reveal that the proposed model outperforms others on the occluded datasets. These comparative experiments demonstrate the effectiveness of the study in addressing the challenge of action recognition in occlusion situations. Our code is publicly available at https://github.com/MJ-Kor/OP-GCN.**

I. INTRODUCTION

In the field of visual perception in robotics, the challenge of seamless human–robot interaction has emerged as a vital problem for research and development. A critical foundation of this interaction is the accurate recognition of human action, a capability that supports robots in understanding and responding intelligently to human actions. In recent years, numerous human action recognition techniques using RGB and skeleton data have emerged, driven by advances in deep learning technology. Among them, skeleton-based action recognition [1]−[3] has received considerable attention due to its potential to recognize human actions by analyzing the skeleton data at a lower computational cost. Moreover, recent studies [4]−[7] have employed graph convolutional networks (GCNs) to extend the use of human skeleton graphs to convolutional layers. Nevertheless, a limitation inherent in existing studies [4]−[7] is the tendency to overlook an occlusion, where parts of the human body become obscured. As illustrated in Fig. 1, a three-dimensional (3D) motion capture camera mounted on a robot [8] or 3D skeleton estimation can extract skeleton data from a person. However, these skeleton extraction systems extract inaccurate skeleton data when human body parts are occluded,

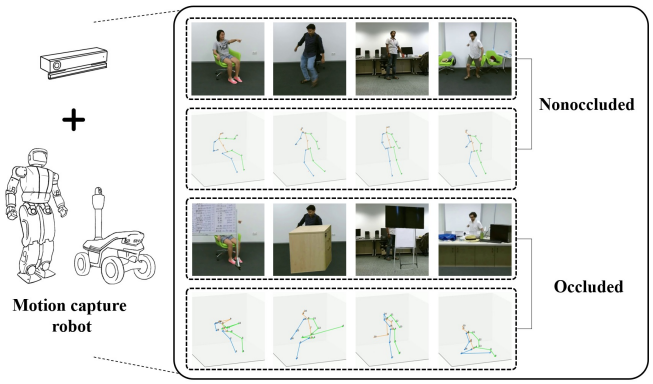

Fig. 1. Results of skeleton extraction system in the nonocclusion situation (second row) and occlusion situation (fourth row).

as if noise has been added. The first and third rows in Fig. 1 present synthetic examples of occlusion, while the second and fourth rows show the results of extracting skeletons using 3D skeleton estimation methods [9], [10]. Because traditional models are sensitive to input data, the data extracted from these obscured views are at a disadvantage compared to the original data. Thus, the influence of occluded skeleton data on accurate action recognition cannot be ignored, and a solution is necessary. We propose a novel and comprehensive approach to address occluded body parts in skeleton-based action recognition to introduce a solution. First, we propose an occluded part detector because existing approaches fail to account for occluded body parts. The proposed occluded part detector is based on an autoencoder trained on typical skeleton data. Moreover, the detector measures the inherent symmetry and angular properties of the human body to effectively detect occluded body parts. By measuring information from these properties, the approach improves the accuracy and robustness of occlusion detection. Second, we propose an occluded part-based group by partitioning these groups based on the five body parts, considering the results of the occluded part detector, and selecting the optimal group. The optimal groups are constructed by excluding the occluded part detector result to retain the meaningful body parts. The optimal group contributes significantly to action recognition in the case of occlusion because, as depicted in Fig. 1, except for the occluded body part, the rest of the body parts are extracted normally.

Finally, we design the occluded part-aware GCN (OP-GCN), including the occluded part-based group and occluded part convolution (OP-Conv). In the occluded part-based group, there may be less correlation between body parts because body parts are constructed independently. To compensate for this, we adopt an interpart association module to improve recognition performance by considering the correlation between parts. Experiments are conducted on datasets characterized by occluded body parts to

This work was supported by the Industrial Fundamental Technology Development Program (No. 20018699) funded by MOTIE of Korea and the IITP grant funded by the Korea government (MSIT) (No.2021-0-02068, RS-2023-00256629, RS-2022-00156287).

The authors are with Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea.{min_hyuk, minju, sbyoo}@jnu.ac.kr

*Corresponding author: Seok Bong Yoo.

validate the effect of the proposed module.

The main contributions are as follows:

- We propose an occluded part detector to detect occluded 3D skeletons using an autoencoder with the joint angles of the human body and the similarity of bone lengths in symmetrical positions.
- We propose an occluded part-based group optimized for occluded body parts to focus on body parts while excluding the occluded body parts.
- We design an interpart association module using a temporal attention map to enhance interaction between parts in a group.

II. RELATED WORK

A. Action Recognition

1) RGB-based Action Recognition: Since the performance of deep learning has dramatically improved, several methods have been proposed for human action recognition based on RGB methods [11]−[14]. Lu et al. [15] proposed a two-stream adaptive weight integration convolutional neural network (AWCNN) with a 3D parallel attention module (PA) called PA-AWCNN. The proposed PA-AWCNN uses the representative integrated feature generated by attention enhancement and feature integration for action recognition. The multiscale vision transformer [16] employs a hierarchical design from the base transformer and a pooling layer to create a multiscale structure. The video swin transformer [17] uses the swin transformer [18] approach on videos, expanding its 2D window concept to a 3D version to capture temporal information for video-related tasks. However, these approaches encounter challenges in accurately identifying human actions because they are sensitive to external factors, such as environmental noise (e.g., background color, light intensity, and attire).

2) RGB and Skeleton-based Action Recognition: Several methods [19]−[21] have emerged that use the simultaneous fusion of RGB and skeleton data to address external factors. Guiyu et al. [19] proposed a neural network that uses a 3D skeleton sequence and a single middle frame from an RGB video as input. Alban Main et al. [20] proposed a modular network combining skeleton and infrared data. The network uses a multilayer perceptron to fuse and exploit RGB and skeleton data. Similarly, Xu et al. [21] proposed bilinear pooling and attention networks, which can effectively fuse multi modality for action recognition. Fusing RGB and skeleton data improves recognition but requires considerable computation and complexity.

3) Skeleton-based Action Recognition: Diverse approaches have been devised for recognizing actions based on skeletons to reduce computation and complexity. Initial approaches [22]−[26] involving manual crafting of features aim to capture joint motion patterns, inputting these features directly into downstream classifiers. With the surge of deep learning, subsequent techniques treat skeleton data as time series, subjecting them to processing using recurrent neural networks [27], [28] and temporal convolutional networks [29], [30]. However, these approaches fail to explicitly model joint relationships, degrading recognition performance. Methods based on GCNs have emerged to address this limitation. As an early GCN-based application in skeleton-based action recognition, spatial-temporal (ST) GCN [31] employs stacked GCN blocks. Each block integrates spatial and temporal modules. Yuxin et al.[32] proposed a novel channelwise topology refinement GCN (CTR-GCN) to dynamically learn topologies and effectively aggregate joint features in various channels. Since the proposal of CTR-GCN, many follow-up studies based on CTR-GCN have been proposed [33]−[35]. However, these CTR-GCN-based models might fail to recognize accurate actions because they do not consider occlusion. Wuzhen et al. [36] proposed an occlusion-aware multistream fusion GCN to consider occlusion. Ioannis et al. [37] proposed a convolutional neural network trained using 2D representations of 3D skeletal motions to deal with occlusion. While these methods consider occlusion, they do not detect occluded body parts, which can be ambiguous in occlusion situations. We propose an occluded part detector and an OP-GCN that uses it to remove this ambiguity.

B. Anomaly Detection

Advances in deep learning have brought several benefits to anomaly detection, enabling features to be extracted at a higher level than traditional methodologies and improving results. These deep learning-based feature extraction methodologies include the autoencoder [38], [39], 3D convolutional Network [40], and 3D convnet autoencoder [41]. In addition, as research on anomaly detection has expanded, recent research has focused on image reconstruction [42]−[47]. This approach detects anomalous events based on the error in reconstructing the input frames of the encoder [48]. A considerable reconstruction error indicates an anomaly, whereas a small error indicates a normal state. This error has also found utility in skeletal video anomaly detection. Thomas et al. [49] employed models based on long short-term memory and the one-dimensional convolutional autoencoder to identify anomalous human behaviors, such as falls. In addition, Onur et al. [50] introduced a method for representing skeleton trajectories that accounted for occlusions, and an autoencoder framework to detect abnormal pedestrian behaviors. Satoshi et al. [51] used a convolutional autoencoder trained on children's skilled gross motor actions. Motion time series images were extracted from video-derived skeletons of kindergarten participants. These images were input into a convolutional autoencoder trained on normal data only. Differences between input and reconstructed pixels were used to detect faulty body movements within abnormal frames. However, these techniques detect abnormal skeletal frames but may fail to determine abnormal skeletal parts. Therefore, we propose a method using additional features from the skeleton to determine abnormal skeletal parts.

III. METHOD

A. Overview

Fig. 2 depicts the overall block diagram of the proposed network. Initially, we used an encoder and decoder trained in the autoencoder format. The abnormal skeleton data are different from the decoder results. This difference is input to the fully connected (FC) layer, concatenated with the two measurement modules to detect the five occluded parts. The result of the FC layer selects the occluded part-based groups whose graphs are constructed with edges and roots. In this case, the occluded part-based groups

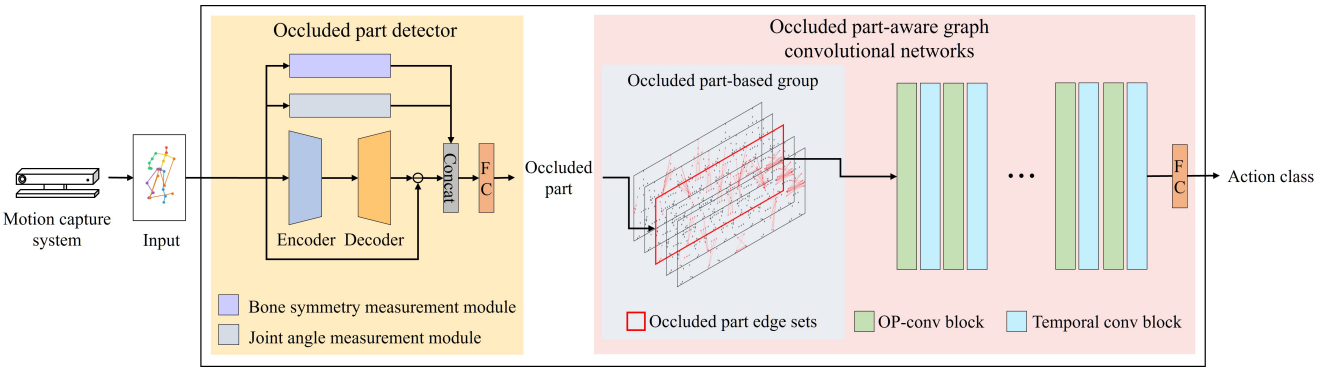

Fig. 2. Architecture of the proposed network for skeleton-based action recognition.

correspond to the occlusion of each body part. The selected group is input into the OP-conv block by creating several adjacency matrices. The OP-conv block generates features using the input adjacency matrix. The generated features are input into the temporal convolutional block. The temporal convolutional block is based on the popular CTR-GCN [32]. After these several OP-conv and temporal convolutional blocks, the FC layer finally recognizes the action. The following sections describe each module in detail.

B. Occluded Part Detector

To detect occluded parts within skeleton data, we used autoencoder-based anomaly detection. The existence of occlusion causes noise in the skeleton data, so the proposed approach is designed to identify abnormal patterns in skeleton data. Fig. 3 depicts the architecture of the proposed occluded part detector. The autoencoder of the occluded part detector is only pretrained on skeleton data with normal patterns and extracts the inherent features of the skeleton in the absence of noise. The pretrained autoencoder tends to reconstruct skeleton data with abnormal patterns similar to skeletons with normal patterns. Consequently, the noise part of the skeleton data has a higher reconstruction error, which can be used as an anomaly score to detect occlusion, as follows:

$$\tilde{x} = x'' - x' = Dec(Enc(x')) - x', \tag{1}$$

where $\tilde{x}$ denotes the reconstruction error, x' represents the input data with linear projection, $Enc(\cdot)$ and $Dec(\cdot)$ indicate the encoder and decoder of the autoencoder, respectively, and x'' denotes reconstruction result.

The human skeleton also has a degree of symmetry in the length of the bones on both sides of the spine, and there are definite limits to the range of motion of each joint. Hence, we implemented the bone symmetry measurement module (BSMM) and joint angle measurement module (JAMM) in Fig. 3 to use the similarity in bone lengths that constitute bilateral symmetry and the angles formed by the joints as supplementary features. First, the equation for the BSMM follows:

$$BSMM_{out} = \Big|\Big|\{u_1, u_2, ..., u_n\}, \tag{2}$$

$$u_i = |b_i^l - b_i^r|, \quad i = 1,2,...,n, \tag{3}$$

$$|b_i^l - b_i^r| = |\sqrt{(s_i^l - e_i^l)^2} - \sqrt{(s_i^r - e_i^r)^2}|, \tag{4}$$

where n denotes the number of symmetrical bone pairs, b_i represents the length of bone, s_i and e_i are the start and end joint nodes of the bone, and l and r denote the left and right sides of the skeleton. In addition, u_i is the absolute difference between b_i^l and b_i^r, representing the similarity of the bone lengths. Note that $\Big|\Big|$ is a concatenation operation. The BSMM uses seven symmetrical pairs of bones: two pairs in the arms, one in the torso, and four in the legs. The equation for the JAMM follows:

$$JAMM_{out} = \Big|\Big|\{\theta_1, \theta_2, ..., \theta_m\}, \tag{5}$$

$$\theta_g = \arccos\left(\frac{\overrightarrow{j_g^\alpha j_g^\beta} \cdot \overrightarrow{j_g^\beta j_g^\gamma}}{\|\overrightarrow{j_g^\alpha j_g^\beta}\| \|\overrightarrow{j_g^\beta j_g^\gamma}\|}\right), \quad g = 1,2,...,m, \tag{6}$$

where m denotes the number of sets with three joint nodes, j_g^α, j_g^β and j_g^γ represent the three joint nodes forming the angle, and $\overrightarrow{j_g^\alpha j_g^\beta}$ and $\overrightarrow{j_g^\beta j_g^\gamma}$ are the vectors originating from j_g^α and j_g^β. In addition, θ_g represents the angle from j_g^α, j_g^β and j_g^γ. Further, JAMM uses two sets for each left and right arm, four for the torso, and three for each left and right leg, resulting in 14 sets of joint nodes. Finally, we use y, the result of concatenating $\tilde{x}$, $BSMM_{out}$, and $JAMM_{out}$, as input to the FC layer to detect the occluded parts.

C. Occluded Part-based Group

This section describes occluded part-based groups. The first step is constructing a graph from the edges and a root tree. To construct a tree from a given skeleton, we must determine a starting node that allows nodes with the same set of occluded part edges to exist in the same semantic space. Furthermore, we determine the set of edges by adopting the five cases that can be the subject of an action as the optimal cases [52]. For example, in Fig. 4, if the result of the occluded part detector is the left leg, we select Group 3 for the left leg. In this case, we start with the end of the right foot because Group 3 is the group that determined that the left foot is occluded. In addition, because the left foot is occluded, there is a probability that the left arm is also not normal. Therefore, we start with the right arm when building the edge set.

Conversely, if the right foot is occluded, we start with the end of the left foot and follow this process. The same applies to the left and right arms. When the torso is occluded, it is constructed from the edge sets, excluding the torso. After selecting the start nodes, the graph is transformed into a root tree, and we define a directed adjacency matrix $\overrightarrow{A_{op}} \in \mathbb{R}^{N_L \times V \times V}$ with N_L hierarchical layers for the N_{op} occluded part edge sets. This process is formulated as follows:

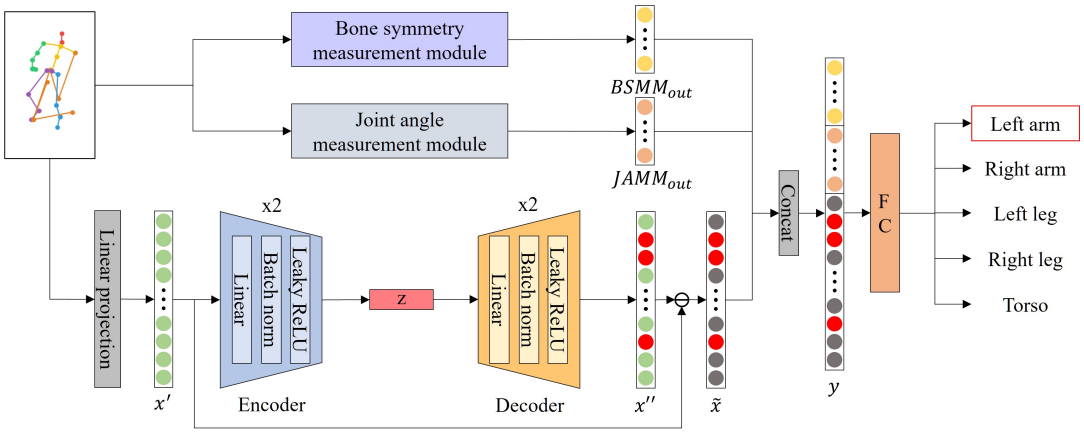

Fig. 3. Architecture of the occluded part detector.

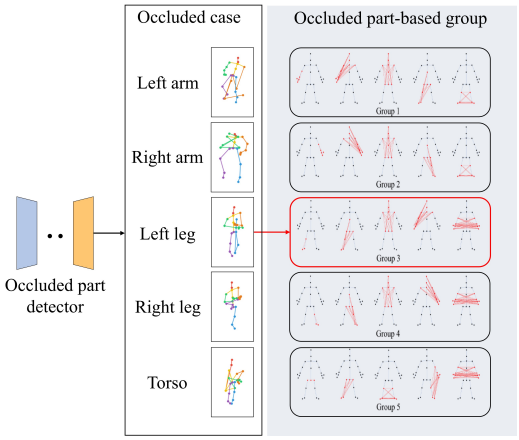

Fig. 4. Architecture of occluded part-based group.

$$\overrightarrow{A_{op}} = \left[\delta(O_1 \to O_2), \cdots, \delta\left(O_{N_{op}-1} \to O_{N_{op}}\right)\right], \tag{7}$$

where O_k represents the kth node group, and $\delta\left(O_k \to O_{k+1}\right)$ represents a set of edges from O_k to O_{k+1}.Further,N_L and N_{op} denote the number of hierarchical layers and occluded part edge sets, respectively, with N_L equal to $N_{op} - 1$, where $\overrightarrow{A_{op}}$ solely encompasses the directed centrifugal edges. For coherence with existing techniques [35], all reverse-directed edges spanning from the leaf nodes of the rooted tree to the start node should be mirrored in the adjacency matrices to encompass the centripetal edges. Moreover, the identity edges of each occluded part-node set must be accounted for to obtain node-specific attributes, leading to the definition of the adjacency matrices $\overleftrightarrow{A_{op}} \in \mathbb{R}^{N_L \times N_S \times V \times V}$ as follows:

$$\overleftrightarrow{A_{op}} = [\delta_1, \delta_2, \cdots, \delta_{N_L}], \tag{8}$$

$$\Delta_k = \Delta\left(\underbrace{O_k \cup O_{k+1}}_{s_{id}}, \underbrace{O_k \to O_{k+1}}_{s_{cp}}, \underbrace{O_{k+1} \to O_k}_{s_{cf}}\right), \tag{9}$$

where Δ_k denotes the set of the three edge subsets of $\Delta(O_k \cup O_{k+1})$, $\Delta(O_k \to O_{k+1})$, and $\Delta(O_{k+1} \to O_k)$ indicate the identity, centripetal, and centrifugal edge subsets, respectively. Through this construction, we create a occlusion-aware skeleton graph edges.

D. Occluded Part Convolution Block

1) Occluded Part Convolution: The representation of 3D temporal skeletal data is denoted as $F_{input} \in R^{3 \times T \times V}$, where V denotes the count of joint nodes, and T represents the temporal window dimension. The OP-conv comprises four parallel branch operations. These operations are the three graph convolutions using an occluded part-based group. In the case of three of these operations, the proposed approach performs a subset wise GCN operation similarly to [31] for each occluded part edge set containing three edge subsets. The output feature map of F_{op} is expressed as follows:

$$F_{op}^k = F_{edge}^k \Big\| _{s \in S} \left\{ \overleftrightarrow{A_s^{(k)}} F_{linear} P_s \right\}, \tag{10}$$

where F_{op} denotes the output feature map of the occluded part convolution and F_{linear} linear denotes a linear projected feature. The operation P_s denotes to pointwise convolution for each edge subsects. GAP in Fig. 5 denotes global average pooling.
where F_{op} denotes the output feature map of the OP-Conv and F_{linear} denotes a linear projected feature. The operation P_s denotes the pointwise convolution for each edge subset. In Fig. 5, GAP denotes global average pooling.

2) Interpart Association Module: Although the occluded part-based group defines more meaningful node relationships than a conventional graph, it may still not be able to extract less correlation between body parts in a group. To enhance correlation, we adopted the interpart association module, formulated as follows:

$$F = \Big\| _{k \in L} F_{op}^k, \tag{11}$$

$$F_{int} = \psi_{global}(F) \cdot \sigma(\psi_{local}(F)), \tag{12}$$

$$\psi_{global}(F) = Sum(F), \tag{13}$$

$$\psi_{local}(F) = P_{local} Max(F), \tag{14}$$

where $\cdot$ and σ denote element-wise multiplication and sigmoid function. In addition, F_{int} represents the integrated feature map with global and local features. Further, ψ_{global} indicates an operation that combines features into a global composition of elements, P_{local} denotes to pointwise convolution, and ψ_{local} denotes an operation for the temporal attention map created by the pointwise

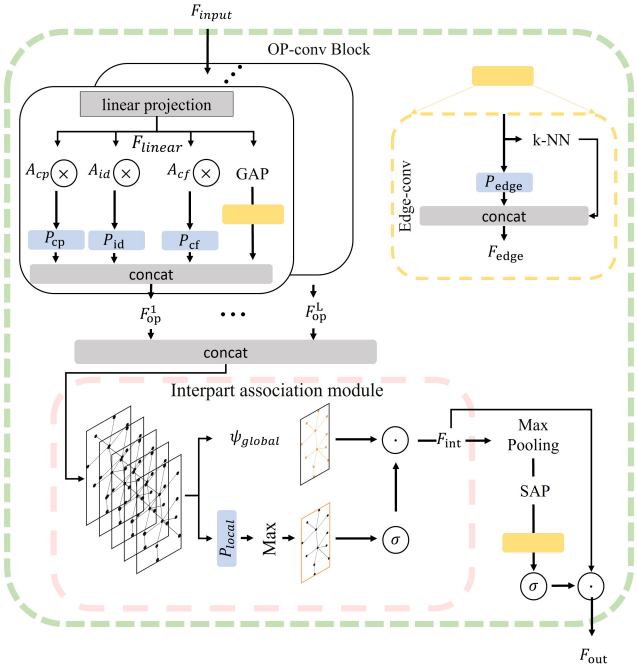

Fig. 5. Architecture of the occluded part-aware convolutional (OP-conv) block.

convolution of the highest value among the features. Finally, the entire process is presented in Fig. 5 and calculated as follows:

$$F_{out} = \sigma(z(\Theta(F_{int}))) \cdot F_{int}, \tag{15}$$

where z and Θ denote the EdgeConv and spatial average pooling (SAP) in Fig.5 adopted in [35]. Moreover, $\times$ denotes the matrix multiplication in Fig. 5. If the occluded part detector result has several parts, we ensemble each part of the recognition results to improve performance.

IV. Experiment

This section presents the experimental results and analysis. We first introduce the experimental setup and datasets for evaluation and training. Finally, we present a discussion and ablation study of the proposed model based on an analysis with other state-of-the-art methods.

A. Data

1) NTU RGB+D dataset: The NTU RGB+D dataset (NTU 60) [53] encompasses 60 distinct action categories and comprises 56 880 action sequences captured using skeletons with 25 joints. These sequences were executed by 40 participants and recorded using three Kinect v2 cameras from diverse angles. The evaluation of this dataset typically employs two specific methodologies in [53].

2) NTU RGB+D 120 dataset: The NTU RGB+D 120 dataset (NTU 120) [54] builds on the NTU RGB+D dataset by incorporating an additional 57 367 sequences characterized by 60 new action classes. The recording involved three camera views and spanned 32 distinct setups, each associated with a specific backdrop and environment. For performance assessment, two specific methodologies were applied, as in [54].

B. Experiment Setting

The experiments were conducted on a single RTX 3090 GPU, and we used [35] as the foundational architecture. The training process involved 100 epochs, with an initial five epochs using a warm-up strategy. We applied the stochastic gradient descent optimizer with a Nesterov momentum value of 0.9 and a weight decay of 0.0004. The learning rate was configured to decay using cosine annealing [55], ranging from a maximum of 0.1 to a minimum of 0.0001 [56]. During the training phase, a batch size of 8 was employed, and we followed the data preprocessing technique outlined in [57]. Moreover, for quantitative results, we created synthetic data using preprocessed data to evaluate for occlusion. When the arm was occluded, we added noise to the joints associated with the arm, and when the leg was occluded, we added noise to the joints associated with the leg. For the arm and leg, we added noise separately for the left and right hand, respectively. We also applied the same process the torso was occluded. For the qualitative analysis, we used data with occlusion from the NTU 120 dataset.

TABLE I
Quantitative Results with other Skeleton-based Action Recognition on the NTU 60 Dataset.

X_Sub										
Occlusion Part	Left Leg		Right Leg		Left Arm		Right Arm		Torso	
Accuracy	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
ST-GCN [31]	42.0	76.4	40.2	74.2	17.5	40.4	2.5	14.9	1.8	13.8
CTR-GCN [32]	50.4	75.0	52.5	77.5	15.4	34.1	5.1	24.5	11.3	32.0
Hyperformer [34]	69.2	89.4	72.5	90.9	52.8	77.7	43.1	68.3	3.2	13.0
HD-GCN [35]	56.8	82.7	44.2	71.4	23.7	45.7	18.5	39.3	10.1	33.3
OP-GCN	**80.0**	**96.5**	**82.2**	**96.1**	**57.2**	**81.4**	**49.9**	**76.3**	**27.6**	**58.1**
X_View										
Occlusion Part	Left Leg		Right Leg		Left Arm		Right Arm		Torso	
Accuracy	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
ST-GCN [31]	63.3	86.8	58.2	79.3	12.7	31.5	2.5	13.8	2.7	15.0
CTR-GCN [32]	52.3	77.7	55.8	80.8	15.8	34.6	8.2	24.3	11.7	32.1
Hyperformer [34]	63.9	83.3	78.2	93.5	42.0	74.4	51.5	73.4	1.9	10.4
HD-GCN [35]	66.8	87.3	73.4	93.8	14.4	33.5	12.9	33.6	8.0	23.3
OP-GCN	**85.3**	**97.5**	**87.6**	**98.0**	**47.8**	**75.5**	**53.2**	**85.3**	**22.5**	**54.7**

TABLE II
Quantitative Results with other Skeleton-based Action Recognition on the NTU 120 Dataset.

X_CSet										
Occlusion Part	Left Leg		Right Leg		Left Arm		Right Arm		Torso	
Accuracy	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
ST-GCN [31]	62.3	80.9	49.0	69.0	14.7	35.6	5.3	11.9	2.0	8.6
CTR-GCN [32]	51.5	79.6	54.9	81.4	12.1	22.4	6.8	16.7	8.4	20.0
Hyperformer [34]	73.4	92.4	75.6	93.4	30.9	57.2	24.8	53.6	0.9	4.7
HD-GCN [35]	63.2	85.8	61.0	84.4	4.2	16.9	3.9	11.0	9.1	26.3
OP-GCN	**76.1**	**93.7**	**76.7**	**94.3**	**49.9**	**73.6**	**31.1**	**57.0**	**9.5**	**26.8**
X_CSub										
Occlusion Part	Left Leg		Right Leg		Left Arm		Right Arm		Torso	
Accuracy	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
ST-GCN [31]	53.7	75.4	62.3	80.9	15.9	38.9	5.4	12.7	2.0	10.3
CTR-GCN [32]	50.7	79.0	51.2	78.9	40.7	20.3	6.8	16.5	5.6	23.5
Hyperformer [34]	58.8	82.8	51.5	77.7	37.7	67.3	27.2	56.5	1.1	5.8
HD-GCN [35]	62.5	86.1	60.9	85.1	2.1	9.6	4.0	12.3	7.9	27.0
OP-GCN	**81.0**	**96.1**	**78.9**	**95.8**	**59.6**	**81.4**	**32.1**	**52.8**	**10.2**	**29.9**

C. Quantitative and Qualitative Results and Analyses

The comparison of the two datasets is presented in Tables I and II. We conducted experiments with these two datasets, consisting of a noisy dataset for five body parts. The performance of each model was measured in terms of the top-1 and top-5 accuracy. Bold in all of the tables indicates the highest score. As indicated in Table I, the proposed model performs recognition robustly on the occluded dataset compared to state-of-the-art methods [31], [32], [34], [35]. Table I and Table II both demonstrate lower accuracy in cases where the torso is occluded. This decline in accuracy can be attributed to the fact that a significant number of the NTU 60 and NTU 120 classes rely on torso movements,

which poses a challenge in accurately identifying the intended action. Meanwhile, in particular, the top-1 accuracy when the torso part is occluded improves by 25.8% for ST-GCN, 16.3% for CTR-GCN, 24.4% for Hyperformer, and 17.5% for HD-GCN on the NTU 60 cross-subject dataset. The proposed model also improves recognition performance for the top-1 and top-5 accuracy in other body parts. Moreover, Table II compares the proposed model with other models obtained from the NTU 120 dataset. In Table II, the proposed model outperforms others regarding the top-1 accuracy average, ST-GCN by 22%, CTR-GCN by 21.92%, Hyperformer by 7.54%, and HD-GCN by 20.38%. This result is because, unlike other models, the proposed model considers and detects occluded body parts and uses them to partialize and select occluded part-based groups to cope with occlusion. Therefore, we experimentally demonstrated that the proposed model is more robust to occlusion situations. Fig. 6 presents the proposed model's action recognition results compared to other models. The first row shows the occlusion situations in the NTU 120. The second row to the last depicted the results of each model. Overall, the proposed model effectively recognizes actions in real-world occlusion situations.

CTR-GCN [32]	Walking towards	Walking towards	Walking towards	Walking towards
Hyperformer [34]	Walking towards	Carry object	Walking towards	Walking towards
HD-GCN [35]	Cheer up	Cheer up	Put on a jacket	Walking towards
OP-GCN	Move heavy objects	Move heavy objects	Move heavy objects	Move heavy objects

Fig. 6. Comparison of action recognition results with various skeleton-based action recognition models in occlusion situations. Blue labels signify the correct action prediction, while red marks mispredictions.

D. Ablation Study

This section presents the experiments that evaluate the performance of the proposed occluded part detector and the effect of the occluded part-based group on action recognition. The dataset used for this study is the NTU 60. Table III lists the accuracy of the occluded part detector when applying each module. Using only the autoencoder results in reduced accuracy. However, combining the autoencoder with BSMM or JAMM increases accuracy by 7.35% and 6.86%, respectively. Additionally, using both modules with the autoencoder improves accuracy by 9.36%. Table IV reveals that the occluded part-based group is listed in the order of left leg, right leg, left arm, right arm, and torso. The highest performance is achieved using the group corresponding to the occluded part. Specifically, when the left arm is occluded, there is an accuracy improvement of 8.43%, 7.85%, and 9.08% over the groups for the left leg, right leg, and torso, respectively. Similarly, when the right arm is occluded, the accuracy improvement is 9.64%, 8.56%, and 14.82% over the same groups. Moreover, Table IV also proves the ablation study for the detectors, as it means when each detector result is wrong. In summary, Table III demonstrates that BSMM and JAMM are effective modules for detecting occluded part in skeleton data. Moreover, the results of Table IV prove that the proposed occluded part-based group method is optimized according to the occluded part.

TABLE III

ABLATION STUDY FOR THE OCCLUDED PART DETECTOR ON THE NTU 60 DATASET.

BSMM	JAMM	Autoencoder	Detection Accuracy
✓	✗	✗	56.35
✗	✓	✗	81.99
✗	✗	✓	86.83
✓	✓	✗	91.49
✓	✗	✓	94.18
✗	✓	✓	93.69
✓	✓	✓	**96.19**

TABLE IV

ABLATION STUDY FOR AN OCCLUDED PART-BASED GROUP ON THE NTU 60 DATASET.

Occluded Part-based Group	Left Leg		Right Leg		Left Arm		Right Arm		Torso	
Accuracy	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
Left Leg	**0.8008**	**0.9652**	0.7998	0.9644	0.7995	0.9644	0.7998	0.9635	0.7985	0.9646
Right Leg	0.8203	0.9600	**0.8228**	**0.9618**	0.8189	0.9604	0.8131	0.9595	0.8203	0.9607
Left Arm	0.4878	0.7354	0.4936	0.7407	**0.5721**	**0.8146**	0.5646	0.8067	0.4813	0.7267
Right Arm	0.4033	0.6895	0.4141	0.6928	0.4743	0.7452	**0.4997**	**0.7683**	0.3515	0.5780
Torso	0.2620	0.5725	0.2679	0.5771	0.2461	0.5467	0.2176	0.5157	**0.2761**	**0.5817**

TABLE V

COMPUTATION COMPLEXITY WITH VARIOUS SKELETON-BASED ACTION RECOGNITION MODELS ON THE NTU 60 DATASET.

Model	FLOPs (G)	Param (M)	FPS
ST-GCN [31]	5.7	3.1	714
CTR-GCN [32]	3.6	1.45	32
Hyperformer [34]	9.62	2.73	250
HD-GCN [35]	3.4	**1.66**	727
OP-GCN	**3.12**	1.76	**833**

E. Complexity

In this section, the experimental results demonstrate that the proposed model is efficient regarding its computational complexity. Table V demonstrates that the proposed model achieved 3.12 giga floating-point operations per second (FLOPs), the lowest of all the models compared. According to Table V, the model does not have the least parameters, but it is still efficient for real-world applications with the highest frames per second (FPS). In conclusion, we prove the efficient complexity of the proposed model for reducing computational cost while maintaining competitive parameter efficiency.

V. CONCLUSION

This paper proposes an occluded part detector and OP-GCN that is robust to occluded body parts. This approach addresses the challenge faced by robots with an obscured human body and surpasses the limitations of conventional models. To solve the challenge, we used an autoencoder trained on normal skeleton data and the angles and symmetry of the human body to detect occluded body parts. The OP-GCN can improve recognition performance by selecting one of the occluded part-based groups from the detected occluded body parts. In conclusion, the model displays advanced human behavior recognition by overcoming the problem of occlusion, where body parts are occluded while improving communication and automation in human-robot interactions.

REFERENCES

[1] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang, "Relational network for skeleton-based action recognition," In *2019 IEEE International conference on multimedia and expo (ICME)*, 2019, pp. 826–831.

[2] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam*, 2016, pp. 816–833.

[3] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017, pp. 4263–4270

[4] L. Hedegaard, N. Heidari, and A. Iosifidis, "Continual spatio-temporal graph convolutional networks," *Pattern Recognition*, vol. 140, p. 109528, 2023.

[5] S. Li, J. Yi, Y. A. Farha, and J. Gall, "Pose refinement graph convolutional network for skeleton-based action recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1028–1035, 2021.

[6] Y. Yoon, J. Yu, and M. Jeon, "Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition," *Applied Intelligence*, pp. 1–15, 2022.

[7] K. Cheng, et al., "Skeleton-based action recognition with shift graph convolutional network," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[8] A. Wang, Y. Makino, and H. Shinoda, "Machine learning-based human-following system: following the predicted position of a walking human," *In 2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4502-4508.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5692-5703.

[10] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7753-7762.

[11] F. Robinson and G. Nejat, "A deep learning human activity recognition framework for socially assistive robots to support reablement of older adults," In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6160–6167.

[12] B. Debnath, et al., "Attentional learn-able pooling for human activity recognition," In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13049–13055.

[13] T. Shen, I. Di Giulio, and M. Howard, "A probabilistic model of activity recognition with loose clothing," In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12659–12664.

[14] Y. Hong, M. J. Kim, I. Lee, and S. B. Yoo, "Fluxformer: flow-guided duplex attention transformer via spatio-temporal clustering for action recognition," *IEEE Robotics and Automation Letters*, vol. 8, no 10. pp. 6411–6418, 2023.

[15] L. Yao, et al., "PA-AWCNN: two-stream parallel attention adaptive weight network for rgb-d action recognition," *In 2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 8741–8747.

[16] Y. Li, et al., "Mvitv2: improved multiscale vision transformers for classification and detection," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[17] Z. Liu, et al., "Video swin transformer," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3202–3111.

[18] Z. Liu, et al., "Swin transformer: hierarchical vision transformer using shifted windows," In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[19] G. Liu, et al., "Action recognition based on 3d skeleton and rgb frame fusion," In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 258–264.

[20] A. M. De Boissiere, and R. Noumeir, "Infrared and 3d skeleton feature fusion for rgb-d action recognition," *IEEE Access 8*, pp. 168297-168308, 2020.

[21] X. Weiyao, et al., "Fusion of skeleton and rgb features for rgb-d human action recognition," *IEEE Sensors Journal*, vol. 21, no. 17, pp. 19157–19164, 2021.

[22] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," In *International conference on machine learning (PMLR)*, 2016, pp. 2014–2023.

[23] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5115–5124.

[24] T. Kipf, E. Fetaya, K. C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," In *International conference on Machine learning (PMLR)*, 2018, pp. 2688–2697.

[25] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, 30, pp.1024–1034, 2017.

[26] D. K. Duvenaud, et al., "Convolutional networks on graphs for learning molecular fingerprints," *Advances in neural information processing systems*, 28, pp.2224–2232, 2015.

[27] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," In *IEEE winter conference on applications of computer vision (WACV)*, 2017, pp. 148-157.

[28] W. Zhu, et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," In *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016, pp. 3697-3703.

[29] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," In *Proceedings of the IEEE conference on computer vision and pattern recognition (ICCV)*, 2017, pp. 3288-3297.

[30] T. Soo Kim, and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20-28.

[31] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018, pp. 7444-7452

[32] Y. Chen, et al., "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.

[33] N. Trivedi, and R. K. Sarvadevabhatla, "Psumnet: Unified modality part streams are all you need for efficient pose-based action recognition," In *European Conference on Computer Vision. Cham: Springer Nature Switzerland*, 2022, pp. 211-227.

[34] Y. Zhou, et al., "Hypergraph transformer for skeleton-based action recognition," *arXiv preprint arXiv:2211.09590*, 2022.

[35] J. Lee, M, Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," *arXiv preprint arXiv:2208.10741*, 2022.

[36] W. Shi, et al., "Occlusion-Aware Graph Neural Networks for Skeleton Action Recognition," *IEEE Transactions on Industrial Informatics*, pp. 10288-10298, 2023.

[37] I. Vernikos, T. Spyropoulos, E. Spyrou and P. Mylonas, "Human activity recognition in the presence of occlusion," Sensors, vol. 23, no. 10, p. 4899, 2023.

[38] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," In Advances in Neural Networks-ISNN 2017: 14th International Symposium, pp. 189–196, 2017.

[39] I. Lee, J. S. Yun, H. H. Kim, Y. Na, and S. B. Yoo, "LatentGaze: Cross-Domain Gaze Estimation through Gaze-Aware Analytic Latent Code Manipulation," *In Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3379-3395.

[40] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 246–255, 2018.

[41] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. S. Hua, "Spatio-temporal autoencoder for video anomaly detection," In *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 1933–1941, 2017.

[42] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, 156, pp. 117–127, 2017.

[43] P. Bergmann, et al., "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *arXiv preprint arXiv:1807.02011*, 2018.

[44] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2626–2634.

[45] D. Gong, et al., "Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection," In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1705–1714.

[46] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 16918–16927.

[47] I. Lee, E. Lee, and S. B. Yoo, "Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition," *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1536-1546.

[48] J. S. Yun, M. H. Kim, H. I. Kim, and S. B. Yoo, "Kernel adaptive memory network for blind video super-resolution," *Expert Systems with Applications*, vol. 238, p. 122252, 2024.

[49] T. Gatt, D. Seychell, and A. Dingli, "Detecting human abnormal behaviour through a video generated model," In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 264–270, 2019.

[50] O. Temuroglu, et al., "Occlusion-aware skeleton trajectory representation for abnormal behavior detection," In *International Workshop on Frontiers of Computer Vision*, pp. 108–121, 2020.

[51] S. Suzuki, Y. Amemiya, and M. Sato, "Skeleton-based visualization of poor body movements in a child's gross-motor assessment using convolutional auto-encoder," In *2021 IEEE International Conference on Mechatronics (ICM)*, 2021, pp. 1–6.

[52] S. H. Han, M. G. Park, J. H. Yoon, J. M. Kang, Y. J. Park, H. G. Jeon, "High-fidelity 3D Human Digitization from Single 2K Resolution Images," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12869-12879.

[53] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[54] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[55] I. Loshchilov, and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[56] J. S. Yun, Y. Na, H. H. Kim, H. I. Kim, and S. B. Yoo "HAZE-Net: High-Frequency Attentive Super-Resolved Gaze Estimation in Low-Resolution Face Images," *In Proceedings of the Asian Conference on Computer Vision,* 2022, pp. 3361-3378.

[57] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1112–1121.