

GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer

Shuaicheng Li^{1*}, Qianggang Cao^{1*}, Lingbo Liu², Kunlin Yang^{1†}, Shinan Liu¹, Jun Hou¹, Shuai Yi¹

¹Sensetime Research, ²The Hong Kong Polytechnic University

{lishuaicheng, caoqianggang, yangkunlin, liushinan, houjun, yishuai}@sensetime.com
liulingbo918@gmail.com

Abstract

Group activity recognition is a crucial yet challenging problem, whose core lies in fully exploring spatial-temporal interactions among individuals and generating reasonable group representations. However, previous methods either model spatial and temporal information separately, or directly aggregate individual features to form group features. To address these issues, we propose a novel group activity recognition network termed GroupFormer. It captures spatial-temporal contextual information jointly to augment the individual and group representations effectively with a clustered spatial-temporal transformer. Specifically, our GroupFormer has three appealing advantages: (1) A tailor-modified Transformer, Clustered Spatial-Temporal Transformer, is proposed to enhance the individual representation and group representation. (2) It models the spatial and temporal dependencies integrally and utilizes decoders to build the bridge between the spatial and temporal information. (3) A clustered attention mechanism is utilized to dynamically divide individuals into multiple clusters for better learning activity-aware semantic representations. Moreover, experimental results show that the proposed framework outperforms state-of-the-art methods on the Volleyball dataset and Collective Activity dataset. Code is available at <https://github.com/xueyee/GroupFormer>

1. Introduction

Group activity recognition is a critical studied problem due to its wide applications in surveillance systems, video analysis, and social behaviors analysis. Different from conventional action recognition, group activity recognition concentrates on understanding the scene of multiple individuals. The intuitive tactic to recognize group activity is to model relevant relations between individuals and infer their

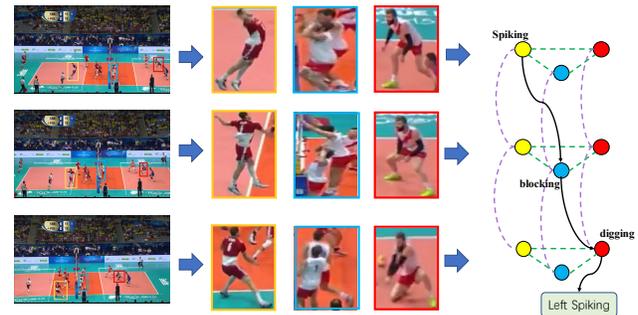


Figure 1. Examples of a clip centered at the annotated frame. The actors with ‘spiking’, ‘blocking’, ‘digging’ actions perform temporally, but they may perform strong spatial-temporal dependencies, which shows the importance of considering spatial and temporal interactions for reasoning about the ‘Left Spiking’ activity.

collective activity. Nevertheless, exploiting individual relations for inferring collective activity is very challenging, especially due to the complicated variations in both spatial and temporal transition in untrimmed scenarios.

Sorts of efforts have been dedicated to capturing relation context in videos for the sake of group activities inferring. Earlier methods [6, 25, 31, 37, 34, 15] utilized recurrent neural network (RNN) to model the dynamics of the individuals, which require a large amount of computational cost. More recent works [20, 47, 18, 33, 23] applied attention-based methods to model the individual relations for inferring group activity. [47] built relational graphs and considered the actor interactions in several frames. [33] captured spatial and temporal self-attention respectively, which are added and utilized to reinforce the mean-field CRF [50]. [20] introduced the standard Transformer encoder as the feature extractor to selectively exploit the spatial actor relations without considering the temporal dynamical information. [23] proposed a relation learning network to model and distill the group-relevant actions and activities by two agents.

However, the aforementioned methods confront two

* indicates equal contribution.

† denotes corresponding author.

challenges that remain to be addressed: 1) build a bridge to model spatial-temporal contextual information integrally and 2) group individuals based on their inter-connected relations for better inferring the global activity context. In the former case, few of the previous methods completely consider the spatial and temporal dependencies in a joint model, while diverse time-series information has strong spatial dependencies, as illustrated in Figure 1. Therefore, capturing spatial-temporal dependencies jointly is critical for reasoning about the group activity. In the latter case, full-connected relations introduced by prior methods [47, 20] is suboptimal since interference information of irrelevant individuals is introduced. Intuitively, not all individuals' relations in the multi-person scenario perform key impacts for the inferring of group activity. As shown in Figure 1, in the volleyball scenario, the interactions between actors with 'spiking' and 'blocking' are considerably higher than the relations between actors with 'spiking' and 'standing', which contribute more for group activity inferring. In other words, the group activity is usually determined by a critical group of individuals with underlying closet relations.

In this paper, we propose an end-to-end trainable framework termed GroupFormer, which utilizes a tailor-modified Transformer to model individual and group representation for group activity recognition. Firstly, we develop a Group Representation Generator to generate an initial group representation by merging the individual context and scene-wide context. Multiple stacked Spatial-Temporal Transformers (STT) are then deployed to augment and refine both the individual and group representation. Specifically, we adopt encoders to embed the spatial and temporal features and apply decoders in a cross manner to build a bridge to model the spatial and temporal contextual information integrally. And a decoder is employed to contextualize the individual representation for augmenting group representation. Besides, in contrast to the existing full-attention manner in Transformer, our STT is further enhanced by a clustered attention mechanism, referred to as Clustered Spatial-Temporal Transformer (CSTT), to model inter-group relations and intra-group relations. More specifically, we dynamically divide all individuals into C clusters, where individuals in the same cluster usually have relevant semantic information. By performing information propagation within each cluster, we can generate compact action features of individuals. Our inter-group attention is to fully model the relations among clusters for facilitating the group activity-aware representation learning. Finally, experimental results show that the proposed network outperforms state-of-the-art methods on the widely adopted Volleyball and Collective Activity datasets. In short, the contributions of this work can be summarized as three folds:

- We propose a new group activity recognition framework, termed GroupFormer, which takes advantage of

query-key mechanism to model spatial-temporal context jointly for group activity inferring.

- A clustered attention mechanism is introduced to assign individuals into groups and build inter- and intra-group relations to enrich the global activity context.
- We perform extensive experiments on the widely adopted Volleyball and Collective datasets. The results show that our GroupFormer outperforms the state-of-the-art methods by a significant margin.

2. Related Works

2.1. Group Activity Recognition

Group activity recognition has been extensively received more attention due to its wide applications. Traditional approaches typically extracted hand-crafted features and then processed by probabilistic graphical models [2, 3, 1, 30, 29, 45] and AND-OR grammar methods [4, 38]. With the rapid development of deep learning, RNN-based methods have achieved remarkable performance due to the learning of temporal context and high-level information [6, 15, 24, 25, 31, 39, 37]. [25] proposed a LSTM model to capture the temporal evolution of each individual and then generate the holistic representation by pooling the actor-leave features. [37] deployed a two-level hierarchy of LSTMs to recognize the group activities more reliably by minimizing the energy of the predictions. In [24], a relational layer is introduced to capture spatial relations for each person for group representation generating. [34] designed a semantic graph and extended it to the temporal dimension by RNN to integrate the spatial and temporal relations.

In recent works, [47] introduced the graph model and built an actor relation graph using GCN to augment the individuals' representation, while the constructed relational graphs are limited to few frames and don't consider the latent temporal individual relationships. [18] adopted self-attention mechanism and graph attention module to model spatial individual dependencies and utilized I3D backbone to capture temporal context. It only captured temporal information using 3D-CNN based backbone and didn't model the temporal relations in their own module. [33] made use of temporal and spatial relations with stacked manner to reinforce the mean-field CRF for learning spatial relations and temporal evolution of the actors. The most related work is [20], which also introduced the transformer to capture semantic representation. However, it only utilized the encoder of vanilla Transformer as a feature extractor for modeling spatial dependencies, and paid much more attention to build strong activity representations using several branches backbone without integrally considering spatial and temporal dependencies. To better exploit the spatial-temporal dependencies integrally, we propose CSTT to not only utilize

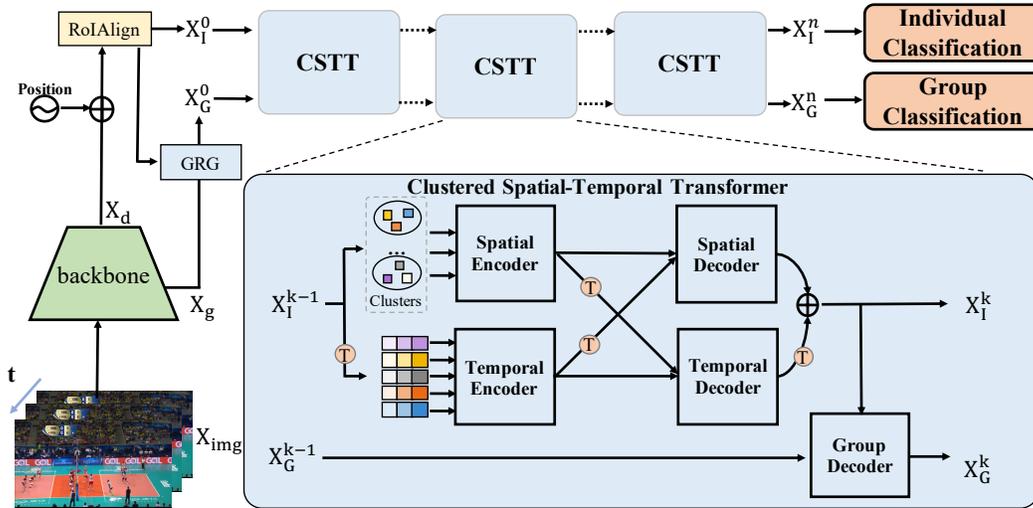


Figure 2. Illustration of our proposed GroupFormer. It contains three main components: 1) a CNN backbone that extracts feature representation of video clips. 2) a Group Representation Generator that initializes the group representation from individual and scene features. 3) a Clusters Spatial-Temporal Transformer that models the spatial-temporal relations and refines the group representation and individual representation.

the encoder to augment semantic representation, but also use a decoder to build a bridge between spatial and temporal relations.

2.2. Transformer

Transformer was first proposed in [43] for sequence-to-sequence machine translation task, and since then has been widely adopted in various natural language processing tasks. The adopted self attention mechanism is particularly suitable for capturing long-term dependencies. Based on Transformer, a series of modifications [16, 13, 36] have been designed to address the limitations in the standard transformer such as computation bottleneck.

Transformer in Vision. The attention mechanism has been widely used in computer vision domain. Due to quadratic cost for the naive application of transformer to images, each pixel attends to all the pixels with the query-key mechanism. Therefore, earlier methods [44, 7] only utilized the self-attention to capture long-range context. [44] designed a non-local attention module to capture long-term dependencies in computer vision tasks. [7] employed a 2D self-attention mechanism to selectively replace 2D convolution and achieve better results than original 2D convolution layer. Very recently, Vision Transformer (ViT) [17], dividing an image into 16×16 patches, took these patches as input and fed them into a standard transformer. The simplicity incurs an extremely high computational price. DETR [10] largely streamlined the detection pipeline based on Transformers, and achieved stronger performances compared with previous CNN based detectors. Unlike ViT [17], DETR adopted CNN based backbone for extracting low-

level features and encoder-decoder based transformer for exploiting high-level concepts.

Spatial-Temporal Transformer. While it has not been explicitly stated, using attention mechanism to model spatial-temporal contexts is quiet general and therefore encompasses many previous works. [48] designed a variant of graph neural network to model the time-varying spatial dependencies. Attention-based spatial-temporal GCN proposed by [21] captured dynamic correlation from the message passing of the graph to learn spatial and temporal features. [12] proposed to employ transformer-based encoders to build spatial and temporal attentions separately and index the temporal attention to the corresponding spatial attention matrix directly. [8] simply extended the ViT [17] design to video by proposing several scalable schemes for space-time self-attention. Previous methods either capture spatial and temporal contexts in stacked manners [8, 48], or utilize parallel modules to extract spatial and temporal features and then fuse them simply [12]. Different from them, we embed spatial and temporal context in parallel, and adopt a decoder to complementary exploit spatial-temporal contextual information.

3. GroupFormer

The proposed GroupFormer is tailored for the group activity recognition involving multiple individuals. The overview of GroupFormer is illustrated in Figure 2. We first process the input image using CNN backbone in Section 3.1 and then initialize the group representation in Section 3.2. Next, in Section 3.3 we introduce our Spatial-Temporal Transformer and clustered attention mechanism in detail.

3.1. Feature Extractor

We adopt Kinetics [26] pretrained inflated 3D network (I3D) [11] as the backbone, and take RGB video clips as input. We start by slicing out a T -frames centered the annotated frame, which is denoted as $\mathbf{X}_{\text{img}} \in \mathbb{R}^{T \times 3 \times H \times W}$ (with 3 color channels). In practice, we extract feature maps from the last convolutional layer and resize to $\mathbf{X}_{\mathbf{g}} \in \mathbb{R}^{T \times C_g \times H' \times W'}$, which can be viewed as the scene feature for the entire video clip. We also generate the higher resolution feature map $\mathbf{X}_{\mathbf{d}}$ from the intermediate *Mixed_4f* layer, and a learnable positional encoding mentioned in [43, 17] is added to the feature map $\mathbf{X}_{\mathbf{d}}$. RoIAlign [22] is applied to extract features for each individuals given N bounding boxes in each video frame. In addition, the pose information of each actor is obtained by AlphaPose [19] and concatenated with the above individual features to provide the final individual features. A fully connected layer is finally adopted to embed the aligned individual features into a D dimensional feature vector for each actor, which can be packed into together named individual representation $\mathbf{X}_{\mathbf{I}} \in \mathbb{R}^{T \times N \times D}$. Besides, we have also conducted experiments on Inception-v3 [41] backbone followed the implementation [47] for a fair comparison with previous methods.

3.2. Group Representation Generator

Group Representation Generator (GRG) is a preprocessed component for initializing group representation in our model. Inspired by [44, 49, 46], video frames can be summarized by a set of feature vectors called visual tokens. Therefore, we transform the scene features and individual features into several tokens respectively and then aggregate them to generate group representation. For scene feature $\mathbf{X}_{\mathbf{g}}$, we view the time dimension as batch dimension and apply a 2D convolution to summarize C_g channels into K . We reshape it into a flattened features $\mathbf{X}'_{\mathbf{g}} \in \mathbb{R}^{T \times K \times (H' \cdot W')}$ and implement a softmax operation to generate the spatial attention matrix $\mathbf{A} \in \mathbb{R}^{T \times K \times (H' \cdot W')}$. Afterwards, we adopt a 2D convolution to embed the scene features into D channels and compute the weighted sum of every pixels with \mathbf{A} to generate K visual scene tokens followed by an Avgpool operation to achieve a scene token shaped as $T \times D$ (we set K to 8). For aligned individual features $\mathbf{X}_{\mathbf{I}}^0$, we feed a learned query shaped as $T \times D$ and individual features into a decoder to generate an individual token. In the end, we fuse the individual token and scene token to form a group token termed initialized group representation $\mathbf{X}_{\mathbf{G}}^0 \in \mathbb{R}^{T \times D}$.

3.3. Clustered Spatial-Temporal Transformer

3.3.1 Canonical Transformer

In this section, we briefly revisit the standard Transformer architecture. The canonical Transformer [43] is proposed

for the sequence-to-sequence task such as language translation. It contains encoders and decoders, both of which apply stacked multi-head attention layers and feed-forward network. The multi-head attention computes the weights by comparing the pairwise similarity between the one and all others, which shows remarkable performance in capturing long-term dependencies. In brief, a sequence of l vectors with the dimension d , represented as $f \in \mathbb{R}^{l \times d}$, is first mapped to the *query*, *key* and *value* using linear projections respectively. Attention matrix is obtained from the scaled dot product of the *query* and *key*. The output features are computed as a weighted sum of the *values* based on the attention weights. Furthermore, The feed-forward network (FFN), consisting of linear transformation and non-linear activation functions, is used to embed the features. More detailed description of the original architecture can be referred to literature [43, 32].

3.3.2 Spatial-Temporal Transformer

Our Spatial-Temporal Transformer (STT) tailored for group activity recognition is designed carefully for augmenting individual representation and group representation. It includes two encoders (a spatial encoder and temporal encoder) in parallel to generate spatial and temporal features respectively. The individual decoders in a cross manner are introduced to decode the spatial-temporal contextual information. Finally, a Group Decoder is applied to augment the group representation. We now explain the proposed model in detail.

Encoders: Despite being localized for each individual, the representation still lacks emphasis on semantic spatial and temporal context, which play an important role in video analysis. Thus two parallel encoders are deployed to embed contextual features. In one branch, we adopt a transformer-based spatial encoder to learn the individual contextual information. Given the input individual representation $\mathbf{X}_{\mathbf{I}} \in \mathbb{R}^{T \times N \times D}$, we view the temporal dimension as the batch dimension and apply an encoder to exploit spatial context for all frames. The process of embedding spatial context for t -frame can be formulated as:

$$\mathbf{Q}^{(t)} = \mathbf{X}_{\mathbf{I}}^{(t)} W_{tq}, \mathbf{K}^{(t)} = \mathbf{X}_{\mathbf{I}}^{(t)} W_{tk}, \mathbf{V}^{(t)} = \mathbf{X}_{\mathbf{I}}^{(t)} W_{tv} \quad (1)$$

$$\mathbf{V}'^{(t)} = \text{softmax}\left(\frac{\mathbf{Q}^{(t)} \mathbf{K}^{(t)T}}{\sqrt{D}}\right) \mathbf{V}^{(t)} + \mathbf{V}^{(t)} \quad (2)$$

$$\mathbf{V}''^{(t)} = \text{FFN}(\mathbf{V}'^{(t)}) \quad (3)$$

where W_{tq}, W_{tk}, W_{tv} are learnable parameters shaped as $D \times D$. $\mathbf{X}_{\mathbf{I}}^{(t)}$ denote the individual feature map in t -th frame. FFN is the feed-forward network in canonical transformer. The feature map of all time steps $\{\mathbf{V}''^{(t)} | t = 1, \dots, T\}$ are packed together into $\mathbf{V}_{\mathbf{s}} \in \mathbb{R}^{T \times N \times D}$.

The other parallel temporal encoder is applied to augment the input features with temporal dynamical evolution clues and enrich the temporal context by highlighting the informative features along time dimension for each individual. The temporal encoder follows the operation of the spatial encoder. The difference with the above spatial encoder is that the temporal encoder views the spatial dimension as a batch dimension. We denote the output temporal dynamic features for n -th individual as $\mathbf{V}^{(n)} \in \mathbb{R}^{T \times D}$, and pack the generated feature maps $\{\mathbf{V}^{(n)} | n = 1, 2, \dots, N\}$ together, represented by $\mathbf{V}_t \in \mathbb{R}^{N \times T \times D}$.

Individual Decoders: Individual Decoders are deployed to consider the spatial and temporal contextual information integrally. Individual Decoders following the standard architecture of Transformer are applied by cross scheme to complementarily exploit spatial-temporal context.

For the spatial decoder, \mathbf{V}_s is viewed as the *actor query*, this N individual queries of each frame are transformed into an output embedding by decoder, where the temporal individual embedding \mathbf{V}_t is regarded as *key* and *value*. The *actor query* captures temporal dynamics from *key* and outputs the updated contextual features. Simultaneously, we also adopt the other temporal decoder by cross scheme. Specifically, the spatial embedding \mathbf{V}_s transposes the time dimension with spatial dimension and can be regarded as the *key* and *value* used by the decoder. The decoder views temporal context \mathbf{V}_t as *time query* and takes it as input and performs feature embedding. During the process, the *key* represents spatial features \mathbf{V}_s along with time order and *time query* of each individual looks up the frames of interest in the video. Finally, the output embedding of these two cross decoders are fused to generate the enhanced individual representation \mathbf{X}_I . The concept of these two decoders with cross manner is to exploit semantic associations based on the spatial context and temporal context for augmenting individual representation.

Group Decoder: Summarizing individuals' interactions in a multi-person scenario is critical to group activity inferring. We introduce a decoder to augment group representation through individual representation. The Group Decoder also follows the pure transformer-based decoder. The difference with the original transformer is that our Group Decoder only contains multi-head cross-attention mechanism and a feed-forward network. It takes the enhanced individual representation \mathbf{X}_I and group representation \mathbf{X}_G as input. Motivated by the learned object query proposed by [10], we adopt group representation, termed as *group query*, to perform group activity context augmenting from individual representation termed as *key*. Thus, the *group query* summarize the overall context from augmented individual representation, and group activity prediction is realized by the updated *group query*. In practice, the output of our Spatial-Temporal Transformer (STT), the enhanced group

representation and individual representation, can be used as input of the latter block. We can stack the STT block repeatedly and learn the underlying semantic representations effectively. Ablation studies are conducted in Section 4.4 to assess the effectiveness of stacked architecture manner.

3.3.3 Clustered Attention Mechanism

Although Spatial-Temporal Transformer (STT) based on fully-connected attention mechanism is capable of modeling individuals' relations, it contains many irrelevant relations. To focus on the crucial group relations, we replace the fully-connected attention with a clustered attention block, called Clustered Spatial-Temporal Transformer (CSTT). It can group the individuals and exploit intra- and inter-group relations for capturing the global activity context.

We first group the individuals into C clusters and then compute two types of attentions: (1) **intra-group attention** as only *query* and *key* from the same cluster are considered. (2) **inter-group attention** as pairwise weighted connection of the clusters are considered. In detail, we define a set of centroid vectors as $\mathbf{M} = (m_1, \dots, m_C) \in \mathbb{R}^{C \times D}$. We utilize mini-batch k-means clustering algorithm to group *query* into C clusters adaptively and update the centroid vectors followed the implementation in [9]. Our intra-group attention is to refine the action-aware feature of each individual by aggregating the information of relative individuals in the same cluster. The inter-group attention is to fully model the relations among clusters for facilitating the group activity-aware representation learning. In detail, the intuitive tactic of building the relations of C clusters is to regard the clusters' centroid vectors as cluster holistic features $\mathbf{M} \in \mathbb{R}^{C \times D}$. We first map the cluster holistic feature into *query*, *key* and *value* using linear functions. Then the inter-attention is obtained by dot production followed row-wise softmax. The cluster features can be calculated by the weight sum of *value*. The updated centroid vector for each cluster can be broadcast to the actors/individuals belonging to the same cluster.

3.4. Network Optimization

Our network is trained in an end-to-end fashion. In our framework, we directly generate the group activity scores \bar{y}_g by group representation obtained from the CSTT. Similarly, another classifier is adopted to predict the individual's action scores \bar{y}_a using individual representation generated by CSTT. For both tasks, we choose the cross-entropy loss to guide the optimizing process:

$$\mathcal{L} = \mathcal{L}_1(y_g, \bar{y}_g) + \lambda \mathcal{L}_2(y_a, \bar{y}_a) \quad (4)$$

where \mathcal{L}_1 and \mathcal{L}_2 denote the cross entropy loss. \bar{y}_g and \bar{y}_a are the group activities scores and individual actions scores while y_g and y_a represent the ground truth labels for the

target group activity and individual actions. λ is the hyperparameter to balance the two terms.

4. Experiments and Analysis

In this section, we experimentally evaluate our proposed network on two widely used datasets. We first introduce these two available group activity datasets, the Volleyball dataset [25] and the Collective dataset [14] in Section 4.1. Then we describe the training details and parameter setting in Section 4.2. In section 4.3, we compare our approach with the state-of-the-art. Finally, a number of ablation studies are conducted to validate the effectiveness of each part within proposed network in section 4.4.

4.1. Dataset

Volleyball Dataset. This dataset [25] contains 55 volleyball videos with 4,830 labeled frames (3493/1337 for training/testing). Each clip is annotated with 8 group activity categories: right set, right spike, right pass, right win-point, left set, left spike, left pass, left win-point. Moreover, the centered frame in each clip is annotated with 9 individual action labels: waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing.

Collective Activity Dataset. This dataset [14] contains 2481 activity clips of 44 video sequences captured by handheld cameras in the street and indoor scenes. Group activities classes are annotated with crossing, waiting, queuing, walking and talking. And the centered frame of each clip are labeled with the bounding boxes of individuals and their individual action classes: NA, crossing, waiting, queuing, walking and talking. The group activity label is assigned as the largest number of the individual actions in the scenes. We follow the same dataset splits as previous works [34, 20].

4.2. Implementation details

For feature extractor, we adopt the Kinetics [26] pre-trained I3D [11] as backbone, separately select 3 frames before and after the annotated frame to be training and testing video clip on both two datasets. An 1×1 convolution reduces the channel dimension of \mathbf{X}_g and \mathbf{X}_d to $D = 256$. The RoIAlign with crop size 7×7 is applied for extracting individuals’ feature map with 256-dimension using the ground truth bounding boxes provided by [6]. For Volleyball dataset, we resize each frame to 720×1280 resolution, for the Collective to 480×720 . For CSTT, we use 1 encoder/decoder layer with 8 attention heads for all encoders and decoders and set dropout probability to 0.1. We choose the number of cluster $C = 4$. Our CSTT is stacked for $\mathbf{b} = 3$ blocks. For the Volleyball dataset we use a batch size of 16 samples and for the Collective dataset we use a batch size of 8 samples. For both datasets, we adopt ADAM [27] to learn the network parameters. Initially, we set the

Method	Flow	Backbone	Group Activity	Individual Action
HDTM[25]		AlexNet	81.9	-
SBGAR[31]	✓	Inception-v3	67.6	-
CERN[37]		VGG16	83.3	69.1
stagNet[35]		VGG16	89.3	-
HRN [24]		VGG19	89.5	-
SSU [6]		Inception-v3	90.6	81.8
ARG [47]		Inception-v3	92.5	83.0
CRM [5]	✓	I3D	93.0	-
Gavrilyuket <i>et al.</i> [20]	✓	I3D	93.0	83.7
Gavrilyuket <i>et al.</i> [20]	✓	I3D+HRnet	94.4	85.9
Ehsanpour <i>et al.</i> [18]	✓	I3D	93.1	83.3
Pramono <i>et al.</i> [33]	✓	I3D	94.1	81.9
Pramono <i>et al.</i> [33]	✓	I3D+Pose+FPN	95.0	83.1
Ours w/o GRG		Inception-v3	93.4	83.2
Ours		Inception-v3	94.1	83.7
Ours	✓	I3D	94.9	84.0
Ours	✓	I3D+Pose	95.7	85.6

Table 1. Comparisons with the state-of-the-art methods on Volleyball dataset in terms of Acc.%. “Flow” denotes additional optical flow input.

learning rate to 0.0001 and decrease by a factor of 10 after 50 and 100 epochs. We set the weight term $\lambda = 1$. Our experiments are all conducted on 8 V100 GPUs.

4.3. Comparison with the State of the Art

We compare our approach with the state-of-the-art methods in two widely adopted datasets, the Volleyball dataset and the Collective dataset separately. For a fair comparison with previous methods, we not only show the results with Pose feature and I3D backbone with RGB and optical flow [40] features but also report the results with Inception-v3 only using RGB features.

Volleyball dataset. The results are listed in Table 1, our method outperforms all of the aforementioned methods with a considerable margin for activity accuracy. To show the effectiveness of our CSTT clearly, we report the experimental result of the proposed model without GRG, which discards GRG and introduces a learned query to be the initialized group representation. It is noting that our model with only RGB features outperforms many previous works [18, 20, 33] though they utilize multi-steps backbone (optical flow, Pose, FPN, Spatial Position), since our GroupFormer exploits both spatial and temporal dependencies integrally. Furthermore, we take pose information as additional input and achieve the best results 95.7% comparing with the previous methods [20, 33]. It’s because that we model the spatial and temporal context in a joint model and augment individual and group representation in parallel. Specifically, compared with [20], the performance of our transformer based method is boosted up by a significant 1.3%, up to 95.7%, indicating that our CSTT can predict the spatio-temporal dynamic of interactions and better enrich group context.

Method	Flow	Backbone	Group Activity
HDTM[25]		AlexNet	81.5
CERN[37]		VGG16	87.2
stagNet[35]		VGG16	87.7
ARG [47]		Inception-v3	91.0
CRM [5]	✓	I3D	85.8
Gavrilyuket <i>et al.</i> [20]	✓	I3D	92.8
Ehsanpour <i>et al.</i> [18]	✓	I3D	89.4
Pramono <i>et al.</i> [33]	✓	I3D	93.9
Pramono <i>et al.</i> [33]	✓	I3D+Pose+FPN	95.2
Ours		Inception-v3	93.6
Ours	✓	I3D	94.7
Ours	✓	I3D+Pose	96.3

Table 2. Comparison with the state-of-the-art methods on Collective dataset in terms of Acc.%. “Flow” denotes additional optical flow input.

Manner	Group Activity	Individual Action
Baseline	91.0	82.1
Spatial only	91.8	82.2
Stacked	92.6	82.8
Parallel	92.2	82.9
Ours	94.1	83.7

Table 3. Ablation study on different variants architectures.

Collective dataset. We further provide detail comparisons with previous methods listed in Table 2 on the Collective dataset. The results with Inception-v3 backbone using RGB features can reach 93.6%, which outperforms most previous methods that use additional optical flow features. Considering the accuracy of 96.3%, a considerable improvement for our model with I3D+Pose backbones is achieved compared to the previous methods [33, 20, 18]. It is worth noting that [20] achieves 92.8% which is 1.9% lower than our result, showing the benefits of spatial-temporal transformer over their spatial transformer based approach.

4.4. Ablation Studies

To validate the effectiveness of the different parts of GroupFormer, we perform ablation studies on the validation set of Volleyball dataset and use group activity accuracy and individual action accuracy as our evaluation metrics. To dispel any concerns that the improvement is simply from additional optical flow and pose information, we only utilize the ImageNet [28] pretrained Inception-v3 [41] as our backbone to extract features from RGB clips in our ablation studies.

Variants of Spatial-Temporal Relations Modeling. To measure the importance of the contextual information gathered by our CSTT, we conduct ablation studies with the following variants. (1) *baseline*: we replace CSTT with a FC layer and a Group Decoder is followed. (2) *spatial manner*: this variant involves a spatial encoder followed by a Group Decoder. (3) *stacked manner*: this variant contains stacked

Clusters	Group Activity	Individual Action
1(w/o cluster)	93.4	83.1
2	93.7	83.6
3	93.8	83.8
4	94.1	83.7
6	93.4	83.5

Table 4. Comparisons of different clusters choices on Volleyball dataset. Cluster set to 1 demonstrate that we adopt original Spatial-Temporal Transformer.

Intra-Attn	Inter-Attn	Group Activity	Individual Action
		93.4	83.1
✓		93.8	83.4
	✓	93.6	83.5
✓	✓	94.1	83.7

Table 5. Comparisons of different cluster attention combinations. Intra-Attn and Inter-Attn denote intra-group attention and inter-group attention respectively.

Blocks	Group Activity	Individual Action
0	91.0	82.1
1	93.6	83.4
2	93.8	83.7
3	94.1	83.7
4	93.9	83.6

Table 6. Comparisons of different setting choices for the number of CSTT blocks.

spatial and temporal encoders, and a Group Decoder. (4) *parallel manner*: this variant consists of parallel encoders, which deal with spatial and temporal context separately and aggregates them by ‘sum’, and then it follows a Group Decoder. Apart from the manner of architecture, the other settings are all the same. Table 3 shows that adopting spatial and temporal in parallel manner improves the performance from 91.0% to 92.2%. Meanwhile, the stacked manner performs slightly better than the models in parallel manner. And our CSTT using Individual Decoders, which build a bridge for jointly exploiting the spatial-temporal contextual information in a cross manner, makes a remarkable performance boost by 3.1% comparing to the baseline. It stresses that learning the spatial-temporal context integrally is effective and important for group activity recognition.

Clustered Attention Manner. Here we investigate the performances of different cluster choices. The results on Volleyball dataset are displayed in Table 4. The STT without clustered attention mechanism can achieve 93.4%, indicating that our proposed model (Spatial-Temporal Transformer) is already capable of learning rich spatial-temporal context for group activity recognition. Meanwhile, we replace the fully-connected attention with clustered attention mechanism called CSTT, leading to a significant boost from 93.4% to 94.1%. It indicates the effectiveness of our clustered attention mechanism. In addition, we gradually mod-

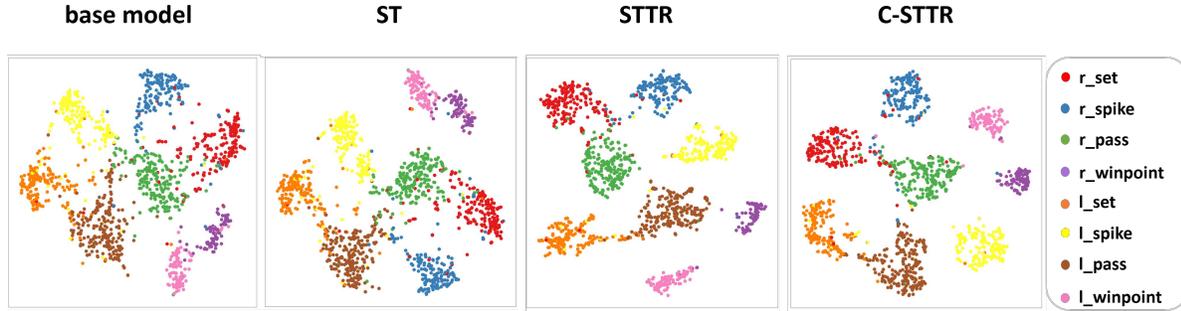


Figure 3. Feature embedding visualizations of the validation set of Volleyball dataset using t-SNE [42] by different model variants. Each clip is visualized as a point and clips belonging to the same group activity have the same color. Best viewed in color.

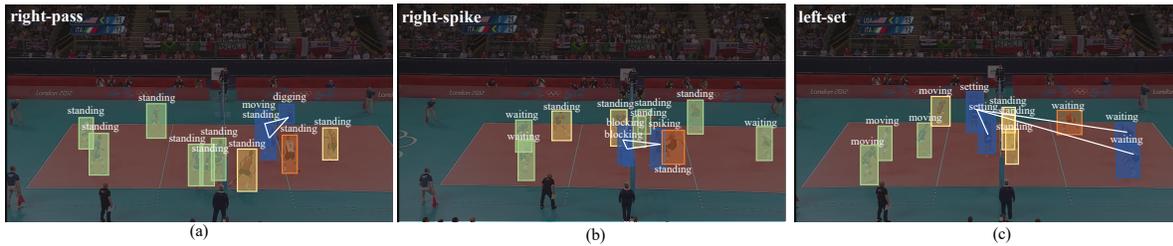


Figure 4. Visualization of some representative clustered individuals. Each video scenario contains the groundtruth labels of activity and individuals. The color of each individual represents the clusters that it belongs to. The white lines represent the close relations in the cluster.

ify the number of clusters and find that the number of cluster set to 4 can reach the best result. Finally, we also evaluate the effectiveness of our two intra-group and inter-group attention. Table 5 reports the detailed results. As expected, we can find that these two attention manner can always contribute to certain performance boosts and the combinations of them reach the best performance.

Investigation of Number of CSTT Blocks. Our clustered Spatial-Transformer can be stacked several blocks to enhance spatial-temporal information. It is essential to evaluate the influence of block number setting. As shown in Table 6, the first column of the table lists the corresponding number of our Clustered Spatial-Transformer. When the number of blocks is set to 0, a FC layer is adopted to replace our CSTT to embed features. We find that a single block outperform baseline by 2.6%, which demonstrate the effectiveness of CSTT. Stacked 3 CSTT blocks reach best results while performance slightly degrades at using 4 blocks.

4.5. Visualization

Feature Embedding Visualization. Figure 3 displays the t-SNE [42] visualization of the video representation learned by our model variants: (1) baseline: replace CSTT with simple FC layers. (2) *Stacked* ST: using stacked spatial temporal transformers. (3) STT: Spatial-Temporal Transformer without Clustered attention mechanism. (4) CSTT: our

Clustered Spatial-Temporal Transformer. In detail, we feed our video representation of the validation set on the Volleyball dataset into 2-dimensional map using t-SNE. We can find that adopting our STT has performed quite well compared with *Stacked* ST. It is noting that our CSTT makes separation better and reaches best results. These visualization results demonstrate that our model is conducive to recognize group activity.

Cluster Visualization. We visualize several examples of the clustered individuals in Figure 4. With clustered attention mechanism, individuals are grouped into sets of nodes, and each set of nodes is closely connected internally. Through exploiting intra- and inter-group relations, it’s easier for the model to capture crucial interaction information and learn activity-aware semantic representations.

5. Conclusion

We propose a new transformer-based architecture termed GroupFormer, which models the spatial-temporal contextual representation for inferring group activity. Furthermore, we introduce a cluster attention mechanism to group the individuals and exploit intra- and inter-group relations together for better group informative features. We perform extensive experiments on two benchmarks. The results show that our GroupFormer outperforms most state-of-the-art methods by a considerable margin.

References

- [1] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hrf: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision*, pages 572–585, 2014. 2
- [2] Mohamed R Amer and Sinisa Todorovic. Sum product networks for activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):800–813, 2015. 2
- [3] Mohamed R Amer, Sinisa Todorovic, Alan Fern, and Song-Chun Zhu. Monte carlo tree search for scheduling activity recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1353–1360, 2013. 2
- [4] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *European Conference on Computer Vision*, pages 187–200, 2012. 2
- [5] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2019. 6, 7
- [6] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4315–4324, 2017. 1, 2, 6
- [7] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019. 3
- [8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 3
- [9] Leon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in neural information processing systems*, pages 585–592, 1995. 5
- [10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020. 3, 5
- [11] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 6
- [12] Kai Chen, Guang Chen, Dan Xu, Lijun Zhang, Yuyao Huang, and Alois Knoll. Nast: Non-autoregressive spatial-temporal transformer for time series forecasting. *arXiv preprint arXiv:2102.05624*, 2021. 3
- [13] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3
- [14] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289, 2009. 6
- [15] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4772–4781, 2016. 1, 2
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4
- [18] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezaatofghi. Joint learning of social groups, individuals action and sub-group activities in videos. *arXiv preprint arXiv:2007.02632*, 2020. 1, 2, 6, 7
- [19] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 4
- [20] Kirill Gavriluk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–848, 2020. 1, 2, 6, 7
- [21] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019. 3
- [22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4
- [23] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 980–989, 2020. 1
- [24] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 721–736, 2018. 2, 6
- [25] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016. 1, 2, 6, 7
- [26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv:1705.06950*, 2017. 4, 6
- [27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

- [28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 7
- [29] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1361, 2012. 2
- [30] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robi-novitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *IEEE transactions on pattern analysis and machine intelligence*, 34(8):1549–1562, 2011. 2
- [31] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2876–2885, 2017. 1, 2, 6
- [32] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064, 2018. 4
- [33] Rizard Renanda Adhi Pramono, Yie Tarng Chen, and Wen Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *European Conference on Computer Vision*, pages 71–90, 2020. 1, 2, 6, 7
- [34] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 1, 2, 6
- [35] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 6, 7
- [36] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. 3
- [37] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. CERN: confidence-energy recurrent network for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5523–5531, 2017. 1, 2, 6, 7
- [38] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4576–4584, 2015. 2
- [39] Xiangbo Shu, Jinhui Tang, Guojun Qi, Wei Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2
- [40] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. 2018. 6
- [41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4, 7
- [42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3, 4
- [44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3, 4
- [45] Zhenhua Wang, Qinfeng Shi, Chunhua Shen, and Anton Van Den Hengel. Bilinear programming for human activity recognition with unknown mrf graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1690–1697, 2013. 2
- [46] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 4
- [47] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019. 1, 2, 4, 6, 7
- [48] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020. 3
- [49] Songyang Zhang, Xuming He, and Shipeng Yan. Latent-gnn: Learning efficient non-local relations for visual recognition. In *International Conference on Machine Learning*, pages 7374–7383, 2019. 4
- [50] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537. IEEE Computer Society, 2015. 1