Selective Kernel and Spatial Grouping Attention Network for Occluded Pedestrian Detection

Yaru Wang
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
wangyarulala@163.com

Yijing Li
School of Foreign Languages
Dalian Maritime University
Dalian, China
katherineli@dlmu.edu.cn

Hua Yu
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
yhiccd@163.com

Qiang Zhang*
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
zhangq@dlut.edu.cn

***Abstract*—Pedestrian detection has achieved significant progress on computer vision tasks in recent years. Most pedestrian detection methods employ deep convolutional neural networks to extract abstract features. However, convolution is a local operation that relies on down-sampling to obtain high-level semantic features, which cannot extract global image information or selectively focus on the input features. Furthermore, since the majority of the pedestrian's body is invisible under severe occlusion, the performance of existing pedestrian detectors remains further improvement. To this end, we propose a novel network with selective kernel and spatial grouping attention, i.e., SKGNet, for the occluded pedestrian detection task. Specifically, we first introduce a lightweight attention module, selective kernel and spatial grouping attention (SKG), which is embedded in the SKGNet's feature extraction backbone. The SKG module combines the properties of the selective kernel (SK) and spatial grouping enhancement (SGE) mechanisms to extract more critical features and improve the expressive ability of feature maps, ultimately improving the detection performance of the network. Moreover, we propose a mask-guided (MG) module to modulate full-body features, which can highlight the visible part of pedestrians while suppressing the occlusion part, thereby significantly improving occlusion detection performance. Extensive experiments show that SKGNet outperforms the existing advanced methods on the CityPersons dataset without excessive extra parameters and computations.**

Keywords—deep learning, convolutional neural network, pedestrian detection, attention mechanism

I. INTRODUCTION

Pedestrian detection plays a critical role in a variety of application scenarios, including intelligent driving assistance, intelligent video surveillance, human-computer interaction, and other fields with high research value. In recent years, deep convolutional neural networks (CNNs) have been widely applied into the pedestrian detection task [1]-[3]. Nevertheless, although the existing pedestrian detectors have made significant progress on benchmark datasets, they still fall well short of human expectations [4].

In general, pedestrian conditions in real-world scenarios are extremely complex, such as background chaos, scale shift and occlusion, which bring great challenges to the pedestrian detection task. Traditional methods typically employ deep neural networks to extract the high-level semantics of objects to detect pedestrians. However, because CNN fails to highlight essential channels and specific spatial positions, this type of method is unable to discriminate the input features. Furthermore, since convolution is a local operation that is applied to a local image to collect local information, the image's global information cannot be extracted, resulting in poor detection results. In recent years, the attention mechanism has been widely applied to pedestrian detection tasks, which enables pedestrian detectors to automatically highlight critical feature channels and spatial locations while extracting features. Therefore, combining attention mechanism with mainstream pedestrian detection framework is an excellent way to overcome the above limitations.

Besides, occlusion is another a tough challenge in practical pedestrian detection applications. Pedestrians are likely to be obscured by other pedestrians or objects in a crowded scene. Despite recent progress on datasets with little or no pedestrian occlusion, detection performance under severe occlusion still needs to be improved. In past years, some approaches [5]-[7] employed an integral detection strategy, training with full-body annotations and assuming the pedestrian was fully visible. However, since the majority of the pedestrian's body is not visible under occlusion, and the detection window incorporates background area, this method reduces the performance of the detection model. Recent approaches [8]-[10] have attempted to deal with occlusion by learning a series of integrated part detectors and jointly training different occlusion modes. However, this kind of method is not conducive to implementation due to its large computation, complex training and reliance on the fusion of partial detection. Different from the above methods, some of the latest methods [11], [12] utilize visible pedestrian information to either regress visible area or learn occlusion patterns, which provides great inspiration to tackle the problem of occlusion detection in our work.

In this paper, we propose an innovative and efficient network with selective kernel and spatial grouping attention, i.e., SKGNet, for the occluded pedestrian detection task. The detection network mainly includes two parts: a feature extractor and a detection head. We propose a lightweight SKG attention module and integrate it into the stacked standard residual blocks of ResNet-101 to generate a more robust and efficient feature extraction backbone SKGNet-101, which

2022 8th International Conference on Virtual Reality (ICVR) | 978-1-6654-7911-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICVR55215.2022.9847953

contributes to enhancement and refinement of the feature map. The SKG aims to enhance specific feature maps based on key attention information, and can automatically adjust the size of the receptive field according to various input information. Besides, it can model a spatial enhancement mechanism and adjust the importance of features by generating an attention factor at each spatial position, enabling the semantic features to autonomously improve learning and expression ability. Furthermore, in order to address the occlusion problem in detection, we propose the Mask-Guided (MG) modulation module. The MG module utilizes the pedestrian's visible body information to generate a pixel-level spatial mask to adjust the multi-channel full-body features, thereby highlighting the visible body part of the pedestrian and suppressing the occluded part. In particular, semantic feature maps of varying levels are merged in the feature extraction stage and successively input to the MG module and detection head for feature modulation and pedestrian detection. The MG module is not restricted to a certain occlusion category and can be easily integrated into a standard pedestrian detector.

In summary, our main contributions are as follows:

- We propose a lightweight attention module SKG, which focuses on the interest area of the pedestrian and improves the expression ability of the feature map, thereby improving the detection performance of the network.
- We propose a Mask-Guided (MG) modulation module for pedestrian occlusion in crowded scenes. The MG module utilizes visible region information to modulate the full-body features, enabling the visible part of pedestrians to be highlighted while the blocked part is suppressed.
- We construct a novel pedestrian detector SKGNet based on attention mechanism and occlusion processing, and conduct extensive experiments on the CityPersons dataset. The results indicate that our SKGNet outperforms its most competitors, e.g., CSP [1], MGAN [2] and Beta-RCNN [13] without generating too many extra parameters and calculations, which are 10.0% MR on Reasonable set, 45.9% MR on Heavy set, 9.3% MR on Partial set, and 5.5% MR on Bare set.

The rest of this paper is arranged as follows. Sec. 2 describes the research background and related work of this paper. Sec. 3 gives details about SKGNet and its modules. Sec. 4 displays the experimental settings, ablation experiments, and results of comparisons with other methods. Sec. 5 reviews the entire work and provides some recommendations for future work.

II. Related Work

A. Pedestrian Detection

Deep learning-based pedestrian detection methods have demonstrated significant benefits and superior performance [14], [15] in recent years. Traditional pedestrian detectors [16] relied heavily on the region proposal and the sliding window to locate and classify candidate objects [16]-[18]. The development of Faster RCNN [19] enabled some two-stage pedestrian detection methods [12], [14], [18] to make progress on the standard benchmark. Recently, a growing number of anchor-free methods emerged. Some single-stage [18], [20] pedestrian detectors, for example, have lately achieved a balance of detection speed and accuracy. Therefore, the current mainstream pedestrian detection algorithms are mainly divided into two categories, anchor-based, and anchor-free methods.

Anchor-based. Anchor-based object detectors have been extensively utilized due to their better accuracy. They entail classifying and regressing anchor boxes with pre-defined scales and aspect ratios. Anchor-based methods fall into two categories: two-stage detectors and one-stage detectors. Among them, the two-stage method strives to improve detection accuracy, nevertheless, the one-stage method will detect much faster than the former. In terms of two-stage pedestrian detection, the RCNN [21] series, which include Fast-RCNN [22], Faster-RCNN [19], and their variants, have emerged as the mainstream structure. The main idea of the two-stage framework is to first generate a series of sparse object candidate proposals through a heuristic method or CNN network (RPN) in the first stage, and then classify and regress them in the second stage. One-stage detectors, on the other hand, are more efficient than two-stage detectors. YOLOv2 [23] and SSD [24] can identify and classify objects simultaneously. One of their major drawbacks is that it is difficult to train, owing to the fact that positive and negative samples are extremely unbalanced, resulting in slightly lower accuracy of the model.

Anchor-free. Anchor-free methods have recently achieved excellent results due to their simple network structure and lack of the need for anchor boxes. They speed up the detection process by directly modeling the detection as a regression task. DenseBox [25] and YOLOv1 [26] are the earliest methods to research anchor-free technology. The former is typically used for face detection, while the latter is typically used for general object detection. Some work, such as [1] and [27], extended the anchor-free idea to pedestrian detection. CornerNet [28] is a new pedestrian detection approach that regards the detection of the object bounding box as a pair of key points. The proposal of the CornerNet propelled anchor-free detection into the key era, which was followed by ExtremeNet [29], CenterNet [30] and so on. In addition, FoveaBox [31] and FSAF [32] employ dense target detection methods.

The network structure of the anchor-free framework is more elegant and simple than traditional anchor-based detectors, achieving the optimal trade-off between speed and accuracy. The key point detection approach, in particular, has lately gained popularity since it is differentiable end-to-end and can directly detect the center point and scale of the object. Therefore, our method essentially employs the anchor-free detection framework based on key points. It is worth noting that the performance of this kind of detector is mainly determined by the expressiveness of the feature map because the key points belong to high-level semantic characteristics. Traditional approaches typically employ down-sampling methods to extract and abstract the feature maps. While our method utilizes the characteristics of the attention mechanism to enhance the expressiveness of feature maps.

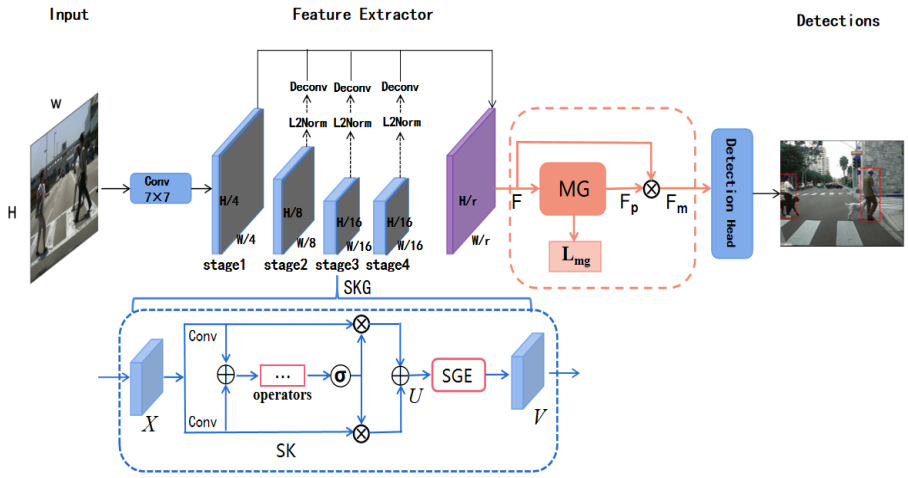


Fig. 1. The overall architecture of our SKGNet, which mainly includes two parts: feature extractor and detection head. The SKG attention module is shown in the blue dashed box, and the Mask-Guided (MG) module is shown in the red box.

B. Attention Mechanism

The attention mechanism has made significant breakthroughs in image processing, natural language processing and other domains in recent years, and it has been proven to be beneficial to enhance model performance. Therefore, it is especially vital to construct neural networks with attention mechanisms. There's been a lot of great work on visual attention in recent years. Wang et al. [33] developed a non-local operation to construct a neural network that can capture long-range dependencies, which was inspired by the classical non-local technique in computer vision. SENet [34] introduced SE block to model the correlation between channels before implementing an attention mechanism on the channels. It can be easily integrated into existing networks at a low cost to improve network performance. STN [35] as a representative of the spatial attention model, proposed the spatial network transformation layer, which enables the model to have spatial invariance. In addition, SKNet [36] implements an attention mechanism on the convolution kernel, allowing the network to select an appropriate convolution kernel autonomously and can adaptively adjust the size of the receptive field according to the input information. Then, SGE [37] proposed spatial grouping enhancement to generate corresponding semantic features in a specific spatial position of the image, enabling each semantic group to enhance its ability to express autonomously. The proposal of SKNet and SGE inspires us greatly in the design and generation of our SKG attention module.

III. Method

In this section, we first present the overall architecture of SKGNet, and then introduce the specific implementations of the SKG attention module as well as the Mask-Guided modulation module respectively.

A. Overall Architecture

Fig. 1 depicts the overall framework of the proposed SKGNet, which consists primarily of two parts: a feature extractor and a detection head. The proposed lightweight SKG attention module is depicted at the bottom in the blue dotted box, which is embedded in standard residual blocks of the backbone network to enhance semantic features. A new Mask-Guided (MG) modulation module is introduced before the detection head, as shown in the red frame in Fig. 1, it can modulate the feature map fused by the backbone.

Feature Extractor. The feature extractor adopts the proposed SKGNet-101 as the backbone network, and the lightweight attention module SKG is embedded in the stacked residual blocks at various stages of the network to extract and enhance effective information. The backbone is divided into four stages, numbered, stage1-4. The feature extractor simply fuses the multi-scale feature image output at each stage into a single feature map for detection, as shown in Fig. 1. For the last three stages, use L2 normalization to adjust the scale of different feature maps, and then use deconvolution operation to standardize their resolution to achieve the purpose of feature fusing.

Take the third stage of the backbone as an example. At this point, define the residual block's output feature map as X, which is fed into the SKG attention module to refine. Firstly, employing the selective kernel mechanism with dynamic kernel selection to get U, followed by the spatial grouping enhancement to obtain V. The refinement process of feature map X can be summarized as follows:

$$U = M_{SK}(X) \tag{1}$$

$$V = M_{SGE}(U) \tag{2}$$

where M_{SK} represents the SK attention mechanism, M_{SGE} is the SGE mechanism, U and V are the feature maps optimized by M_{SK} and M_{SGE} successively.

The feature map extracted and fused by the backbone is fed into the Mask-Guided module, which modulates the whole input body features with visible boundary box information and outputs a pixel-level feature map, as depicted in the red dotted line box in Fig. 1. Aiming to emphasize the visible part of the pedestrian's body while suppressing the obscured parts.

Detection Head. The feature map modulated by the MG module is fed to the attached detection head to obtain the

pedestrian detection result. The detection head, similar to [4], consists of a 3×3 convolution layer, followed by three 1×1 convolution prediction layers for center point position prediction, offset prediction, and scale regression.

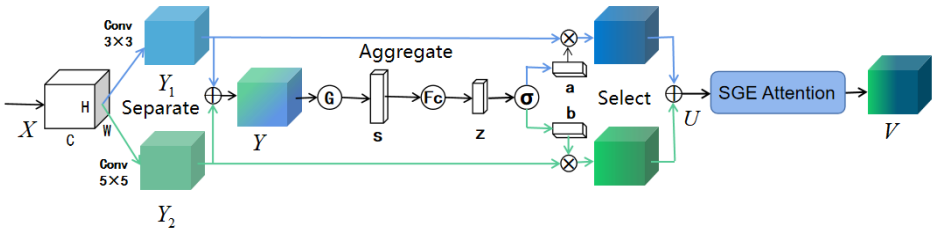


Fig. 2. Detail display of SKG attention module, which includes four steps: Separate, Aggregate, Select and Spatial grouping enhance.

B. SKG Attention Module

It can be seen from the above that in the feature extraction stage of the network, the standard residual block of the backbone is combined with the lightweight SKG module to build an efficient feature extraction block, which is conducive to extracting more significant detection information. The lightweight SKG module can improve network detection performance with a small number of additional parameters and calculations and can also be integrated into other mainstream backbone networks.

Although we are all aware that the size of the receptive field of neurons can be adjusted in response to external stimuli, this is rarely exploited in the actual construction of CNN. Based on this, [36] proposed a dynamic kernel selection mechanism capable of automatically selecting multiple different kernel sizes, in which each neuron can adjust its receptive field size adaptively based on varied input scales. And it can enable the network to capture features more accurately. Furthermore, [37] proposed to generate corresponding semantic features in a specific spatial position of the image. It models a spatial enhancement mechanism in each feature group and adjusts the importance of each sub-feature by generating an attention factor. The attention factor can suppress the noise in the process and activate the specific semantic feature area, enabling each semantic group to enhance its ability to express autonomously.

Therefore, we propose the SKG attention module, which combines selective kernel (SK) and spatial grouping enhancement (SGE) attention mechanisms to construct an effective feature extraction block. SKG adjusts the receptive field of neurons first with dynamic kernel selection, then groups channels and enhances spatial attention with SGE. This method of combination leverages the benefits of SK and SGE, allowing the network to extract more critical feature information and enhance detection performance. Subsequent ablation experiments also confirmed it. The entire SKG module implementation process can be divided into four steps: *Separate*, *Aggregate*, *Select* and *Spatial grouping enhance*, as illustrated in Fig. 2.

Separate: Assuming that the previous block's output feature map is $X \in \mathbb{R}^{H \times W \times C}$, it is first sent to the two convolution layers with the convolution kernel sizes of 3×3 and 5×5 for processing, and generate two feature maps $Y_1 \in \mathbb{R}^{H \times W \times C}$ and $Y_2 \in \mathbb{R}^{H \times W \times C}$with different semantic information. Both convolution layers include three operators in sequence: grouped convolutions, switchable normalization, and the ReLU function. It is worth noting that in training, we utilize the dilated convolution of 3×3 kernel with an expansion scale of 2 rather than the traditional 5×5 kernel convolution to improve efficiency even further. The following is the Y_1 and Y_2 generation process:

$$Y_1 = Conv3 \times 3(\delta(\mathcal{SN}(X))) \tag{3}$$

$$Y_2 = Conv5 \times 5(\delta(\mathcal{SN}(X))) \tag{4}$$

where $\mathcal{SN}$, δ and $Conv$ represent Switchable Normalization, ReLU activation function and convolution, respectively.

Aggregate: As mentioned above, the purpose of the SKG module is to enable neurons to adjust their receptive field size autonomously according to external information. Therefore, the information from all branches needs to be integrated to achieve this goal. Firstly, we simply fuse multiple branches Y_1 and Y_2 via element summation operation to generate feature map Y. Then employ global average pooling to generate channel statistical features $s \in \mathbb{R}^C$ to integrate the global information of Y. Then, a basic 1×1 fully connected layer is employed to reduce the dimension and generate a compact feature $z \in \mathbb{R}^D$, which improves the accuracy and efficiency of the adaptive selection. The formula for the Aggregate process is as follows:

$$z = \mathcal{F}_{fc}(\mathcal{F}_{gp}(Y_1 + Y_2)) \tag{5}$$

where $\mathcal{F}_{gp}$ and $\mathcal{F}_{fc}$ denote the Global average pooling and full connection layer, respectively.

Select: This operation primarily applies the softmax operator to the compact feature z, so as to adaptively select information of different spatial scales, and the calculated attention weight vectors are denoted as a and b, respectively. Then applying the two attention weights a and b to weigh and sum up the feature maps Y_1 and Y_2, which were previously split by different convolution kernels. Finally, we obtain the feature map $U \in \mathbb{R}^{H \times W \times C}$. The entire procedure is depicted in the following formula:

$$[a, b] = softmax(z) \tag{6}$$

$$U = a \cdot Y_1 + b \cdot Y_2 \tag{7}$$

Spatial grouping enhance: SGE aims to improve the learning of different semantic sub-features within each group and intentionally self-enhance their spatial distribution in the group, so that each group of features is spatially robust and well distributed. All SGE enhancements are operated within the group, requiring almost no additional parameters and calculations. The attention enhancement formula is summarized as follows:

$$V = SGE(U) \tag{8}$$

Firstly, the attention weight feature map $U \in \mathbb{R}^{H\times W\times C}$ output by the previous SK attention mechanism is divided into G groups along the channel dimension. Here we take one set of $U' \in \mathbb{R}^{H\times W\times \frac{C}{G}}$ as an example to explain. The distribution of feature maps is always affected by inevitable noise and similarity between features. Therefore, in order to obtain the feature map with ideal distribution, we can utilize the global statistics of the entire group space to enhance the learning of specific semantics. Specifically, the global average pooling function $\mathcal{F}_{gp}(\cdot)$ is utilized to approximately calculate the semantic feature vector learned by the group. Then, a dot product calculation is performed for global statistical features and local features to measure the similarity between them. Then, in order to avoid coefficient divergence between distinct samples, we normalize the dot product result and obtain the normalized coefficient c. Finally, the normalized coefficient spatially scales the original feature U' with the sigmoid function $\sigma(\cdot)$ to obtain the final enhanced feature group $V' \in \mathbb{R}^{H\times W\times \frac{C}{G}}$. The overall enhancement process can be summarized as follows:

$$c = \mathcal{N}\left(U' \cdot \mathcal{F}_{gp}(U')\right) \tag{9}$$

$$V' = U' \cdot \sigma(c) \tag{10}$$

where $\mathcal{F}_{gp}$, $\mathcal{N}$ and σ represent Global average pooling, Normalization and sigmoid activation function, respectively.

C. Mask-Guided Module

Occlusion is a common and tricky challenge in pedestrian detection. Reference [2] proposed Mask-Guided Attention Network to improve the performance of the anchor-based two-stage pedestrian detector. In order to make our anchor-free pedestrian detector more conducive to detecting occluded pedestrians, we propose a mask-guided modulation module based on [2], which is highlighted in Fig. 1 with a red box. Different from the MGA branch in [2], which uses Rectified Linear Unit (ReLU) twice in succession to extract features, our MG module utilizes Leaky ReLU (LReLU) function to replace the original second ReLU, alleviating the possible dead RELU problem and improving computational efficiency. Subsequent ablation experiment also confirmed the effectiveness of our improvement. The MG module generates a spatial attention mask based on pedestrian visible parts, which can be used to modulate the fusion features. The features modified by the MG module can assist the subsequent detection head in detecting partly or severely occluded pedestrians with greater confidence.

Fig. 3 shows the MG structure. The input of the module is the merge of feature maps with various resolutions, and the output is a feature map modulated by the spatial mask. The mask-guided module is described in detail below.

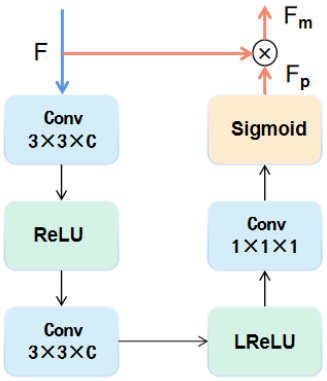
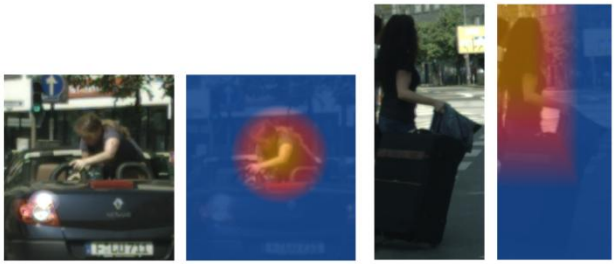


Fig. 3. The detailed content of the Mask-Guided modulation module.



(a) (b) (c) (d)

Fig. 4. Visualization of modulation features. (a) and (c) are the original pedestrian images to be detected. (b) and (d) are the feature maps after MG module modulation.

Assume that the feature map fused in the previous stage is $F \in \mathbb{R}^{H\times W\times C}$. We set H and W to 7 and C to 768 in the experiment. F is first fed into the two sequential combinations of 3×3 kernel size convolution layer and activation function to extract features, the activation functions are ReLU and LReLU respectively. Next, through a 1×1 kernel convolution layer and a sigmoid function to generate a pedestrian probability map $F_p \in \mathbb{R}^{H\times W\times 1}$. Similar to [2], the probability map F_p then modulates the input multi-channel feature F by element dot product with the feature F_i of each channel in F, and generate the weighted feature F_m. The formula is as follows:

$$F_P = \delta_S\ Conv1\left(\delta_{LR}\ Conv3\ (\delta_R\ Conv3\ (F))\right) \tag{11}$$

$$F_{m_i} = F_i \odot F_p, i = 1,2,\dots,C \tag{12}$$

where δ_R represents ReLU activation function, δ_S represents the sigmoid function, i indicates the channel and $\odot$ denotes the element dot product. Finally, the modulated feature F_m is fed into the detection head for pedestrian detection.

The loss function L_{mg} of the MG modulation module defined by binary cross-entropy loss, refer to [2], is calculated as follows:

$$L_{mg} = BCELoss\left(y_i(x,y), {y'}_i(x,y)\right) \quad (13)$$

where ${y'}_i(x,y)$ is the predicted result of the MG module, and $y_i(x,y)$ indicates the ground truth.

As shown in Fig. 4, the features modulated by the MG module highlight the visible part and suppress the occluded part, making it easier to detect occluded pedestrians.

TABLE I. STRUCTURE DETAILS OF BACKBONE NETWORK SKGNET-101

Layer name	Output size	Architecture of SKGNet-101
Conv1	112×112	7 ×7, 64, stride 2
Conv2_x	56×56	3×3 max pool, stride 2 1×1,64 SKG module ×3 1×1,256
Conv3_x	28×28	1×1,128 SKG module ×4 1×1,512
Conv4_x	14×14	1×1,256 SKG module ×23 1×1,1024
Conv5_x	7×7	1×1,512 SKG module ×3 1×1,2048
	1×1	7×7 global average pool, 1000-d fc, softmax

D. Details of Backbone Network

Table I shows the overall architecture details of SKGNet-101, which is the backbone network of the proposed SKGNet. The designed SKG attention modules are embedded in the standard residual blocks of ResNet-101 to form SKG blocks, and multiple SKG blocks are stacked to construct the backbone network SKGNet-101. SKGNet-101 introduces only a few parameters and calculations compared to the original ResNet-101 because the SKG module is a lightweight independent module. As shown in Table I, SKGNet-101 mainly consists of 4 stages, each of which is stacked with 3, 4, 23 and 3 SKG residual blocks respectively. Each SKG block is made up of convolution layers of varied kernel sizes and an SKG attention module. The number of SKG residual blocks at each stage is shown outside brackets.

IV. EXPERIMENTS AND ANALYSIS

In this section, we will first introduce the pedestrian detection benchmark, evaluation metric and implementation details individually. Then, we conduct ablation experiments to evaluate the efficacy of our proposed pedestrian detector. Finally, we compare the proposed pedestrian detector with the state-of-the-art pedestrian detection model on the CityPersons dataset and report the comparison results.

TABLE II. COMPARISON OF METHODS USING DIFFERENT ATTENTION MODULES ON THE CITYPERSONS DATASET

Method (without MG)	#Parameters (MB)	Test Time (ms/img)	Reasonable MR^{-2}(%)
ResNet-101(*SN*)	44.55	259.07	11.71
SKNet-101	45.68	266.05	10.74
SGENet-101	44.55	265.96	10.50
SKGNet-101(ours)	45.68	264.20	**10.19**

TABLE III. COMPARISON OF RESULTS OF SK AND SGE IN DIFFERENT PERMUTATIONS

Arrangement (without MG)	#Parameters (MB)	Test Time (ms/img)	Reasonable MR^{-2}(%)
SK+SGE	45.68	264.20	**10.19**
SK//SGE	45.68	263.89	10.32
SGE+SK	45.68	273.14	10.41

TABLE IV. COMPARISON OF PEDESTRIAN DETECTOR SKGNET WITH AND WITHOUT MG MODULE

Method	#Parameters (MB)	Test Time (ms/img)	Heavy MR^{-2}(%)
SKGNet (without MG)	45.68	264.20	48.63
SKGNet (with MGA)	45.68	270.05	46.29
SKGNet (with our-MG)	45.68	270.10	**45.94**

A. Datasets and Evaluation Metrics

Datasets. We conduct our experiments on one of the largest pedestrian detection benchmarks, CityPersons [38], to verify the efficacy of the proposed method. CityPersons is a more challenging large-scale dataset with high resolution for pedestrian detection. It consists of 5,000 images (2,975 for training, 500 for validation and 1,525 for testing) which provide pedestrian box annotations for full body and visible region. CityPersons contains four subsets with various degrees of occlusion: Bare, Reasonable, Partial, and Heavy, with visible pedestrian body ratios of [0.9, 1], [0.65, 1], [0.65, 0.9], [0.2, 0.65]. The overall dataset has a density of roughly 7 pedestrians per image. We train our model on the official training set and test it on the validation set.

Evaluation Metric. We employ the most common and standard pedestrian detection evaluation metric to report the detection results, log-average Miss Rate(denoted as MR^{-2}), which is computed over the False Positive Per Image (FPPI) range of $[10^{-2}, 10^{0}]$ [39]. And the lower the value, the better the detection performance.

B. Implementation Details

We implement the proposed model on Pytorch. In particular, in order to optimize the network, we adopt the Adam solver to achieve better and more stable performance. The detector's original backbone is ResNet-101 pre-trained on ImageNet [40]. We train the network on two Nvidia Tesla V100 GPUs during the experiment. The resolution of the input image is fixed at 640 × 1280 pixels based on the computing capabilities of the hardware device and GPU. In addition, for the CityPersons dataset, we set up a mini-batch of 4 images, with each GPU allocating 2 images. The learning rate is 2×10^{-4} and training is finished after 150 epochs (744 iterations per epoch).

We employ the anchor-free detector Center and Scale Prediction(CSP) proposed by [1] as the basic detection framework. The structure of this kind of one-stage detector is simple, but the detection accuracy is unsatisfactory. In order to make the detection architecture more robust and accurate, we made the following improvements:

1) Backbone network: We employ ResNet-101 as the backbone to improve detection performance. The third part of ResNet-101 has 23 bottleneck blocks, which is 51 layers more than ResNet-50's 6 blocks. This enables the network to extract higher-level semantic information without altering location information.

2) Normalization layer: To solve the defect that the Batch Normalization (*BN*) layer is susceptible to batch size, we replaced the original *BN* layers in the experiment with Switchable Normalization (*SN*) layers. This enables the detector to fully exploit high-level semantic information [41] in the feature map to improve performance.

3) Loss function: The total detection loss of our SKGNet is made up of three components: center classification loss, scale regression loss and offset prediction loss. Specifically, Focal Loss is employed as the classification loss, vanilla L1 as the scale loss and Smooth L1 as the offset loss. The weights of the three items are adjusted to 0.01, 0.05 and 0.1.

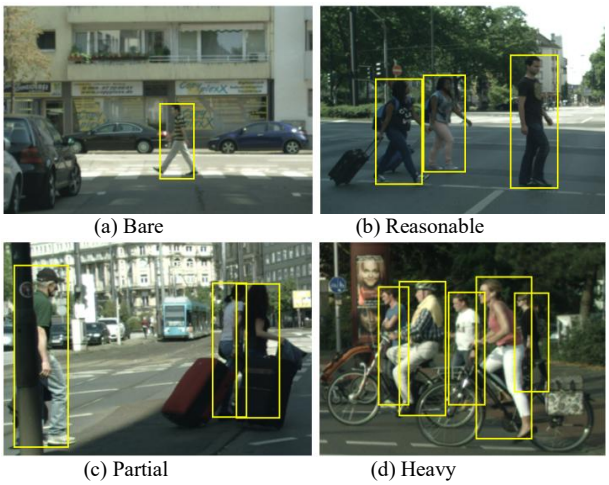
C. Ablation Study

In this section, we conduct ablation experiments on the CityPersons [38] dataset to verify the efficacy of the various parts of our pedestrian detection model. The evaluation experiment consists primarily of the three items listed below:

The significance of the SKG attention module. The experiments in this section primarily explain and validate the importance of the SKG attention module. The backbone network employed in the experiment is our improved ResNet-101, and none of the four methods utilized MG modules for occlusion processing. The parameters, test time, and experimental results of the four different methods are compared in Table II. #Parameters column specifies the number of parameters for different network models. Test Time denotes the duration of a single image with a resolution of 640 × 1280. The evaluation index MRs is calculated based on the Reasonable subset for a more intuitive comparison, and the best result is bolded.

The results in Table II show that, as compared to the original backbone network, embedding the attention module in the stacked residual blocks can indeed improve the model's performance to varying degrees, with only a few parameters added. Furthermore, when compared to embedding the SGE and SK modules separately, combining them as the SKG attention module integrated into the network can fully exploit the enhancement benefits of the two attention mechanisms, resulting in the best experimental result of 10.19%. This is because the SKG module not only possesses the adaptive kernel selection strategy of the SK attention mechanism, but it can also utilize the spatial grouping of SGE to enhance the network's semantic feature learning ability, and the function is more powerful. As a result, the SKG module is adopted as the core feature extraction module in our proposed pedestrian detector.

The optimal way to combine SK and SGE. In this section, we compare three different setups of the two attention mechanisms SK and SGE in the SKG attention module, namely SK+SGE, SK//SGE, and SGE+SK respectively, with the same experimental settings as above. SK+SGE means that they are connected in the order of SK before SGE, SK//SGE implies that they are connected in parallel, and SGE+SK means that they are connected in the order of SGE before SK. Table III shows that, with the same network parameters and a tiny difference in test time, all three connections outperform the original network in terms of detection results. And the setup with the best performance is based on the SK+SGE arrangement, followed by SK//SGE, and finally SGE+SK. As a result, the first arrangement is adopted as the SKG module combination mode in our pedestrian detector.



(a) Bare (b) Reasonable

(c) Partial (d) Heavy

Fig. 5. Detection examples of our SKGNet on the CityPersons dataset.

The impact of the Mask-Guided module. Comparative experiments in this section have fully demonstrated the efficiency of the MG modulation module for occlusion detection. In order to further validate the module's effectiveness, the evaluation index MRs is calculated based on the heavy set, and the best result is bolded. The other comparison metrics are the same as those listed in the above section. The experimental results in the Table IV show that the use of MG effectively reduces the MR^{-2} on the Heavy occlusion set from 48.63% to 45.94%, and the performance is

improved by 2.69% without requiring a notable increase in extra parameters or test time. In addition, compared with the direct use of MGA branch [2], our improved MG module has better detection result, which also reflects the necessity of activation function replacement. Therefore, it is necessary to further process and optimize the feature map prior to the detection head with the MG modulation module, which can significantly improve the problem of pedestrian occlusion.

D. Comparison to the State-of-the-Art

In this part, we extensively compare the proposed SKGNet model with the state-of-the-art approaches on the validation set of CityPersons benchmarks. The results are presented in Table V. Furthermore, to illustrate the effect of our detector more intuitively, we display the detection results of the proposed SKGNet detector on the CityPersons dataset in Fig. 5. The detection samples on the four subsets of CityPersons, which are bare, reasonable, partial, and heavy, are shown in Fig. 5 from left to right and top to bottom.

The hardware in Table V refers to the type of GPU that was used to train the network in the experiment, while the backbone is employed by various detection models to extract features. The evaluation metric MRs is reported on four sets of CityPersons with varying levels of occlusion. The best results are shown in bold. As shown in Table V, the proposed SKGNet outperforms other state-of-the-art approaches on the most heavily occluded Heavy set and the most sparse Bare set. And compared with the two-stage detection method of MGAN that also adopts occlusion processing, the performance of our detector is better, which not only performs occlusion processing but also adopts attention mechanisms for feature extraction. On the Reasonable set and Partial set with moderate occlusion, our method ranks second in performance. The competitor APD beats our detector on both subsets due to the use of a more powerful backbone network, DLA-34, and additional post-processing. With ResNet-50 as its backbone, its MR in the two subsets would increase to 10.6% and 9.5%, respectively. In addition, due to the limitations of hardware device and GPU computing capacity, the image resolution of our method was uniformly set at 640 × 1280 during training. Compared to detectors with a resolution of 1024 × 2048, the performance of our detector is somewhat limited, as higher resolutions will generally yield better detection results.

V. CONCLUSION

In this paper, we propose SKGNet, an innovative and efficient pedestrian detector based on attention mechanism and occlusion processing. In order to enable the network to extract more critical features, we propose a lightweight attention module SKG, which combines the SK's automatic kernel selection with the SGE's spatial grouping enhancement. The SKG module can improve network performance by enhancing the expressive ability of the feature map. In addition, we propose a mask-guided (MG) modulation module for pedestrian detection in crowded scenes. This module employs visible area information to modulate the full-body features, significantly improving occluded pedestrian detection performance. The results of the comparison and ablation experiments on the CityPersons dataset indicate that our proposed SKGNet pedestrian detector outperforms the competitors.

TABLE V. COMPARISON OF OUR SKGNET AND OTHER STATE-OF-THE-ART DETECTION METHODS ON THE CITYPERSONS VALIDATION SET

Method	Hardware	Backbone	Reasonable	Heavy	Partial	Bare
FRCNN [38]	GPU	VGG-16	15.4%	-	-	-
RepLoss [14]	GPU	ResNet-50	13.2%	56.9%	16.8%	7.6%
OR-CNN [42]	Titan X GPU	VGG-16	12.8%	55.7%	15.3%	6.7%
ALFNet [43]	GTX1080Ti GPU	ResNet-50	12.0%	51.9%	11.4%	8.4%
Adaptive NMS [18]	Titan X GPU	ResNet-50	11.9%	54.0%	11.4%	6.2%
CSP [1]	GTX1080Ti GPU	ResNet-50	11.0%	49.3%	10.4%	7.3%
MGAN [2]	NVIDIA GPU	VGG-16	10.5%	47.2%	-	-
APD [3]	GTX1080Ti GPU	DLA-34 ResNet-50	8.8% 10.6%	46.6% 49.8%	8.3% 9.5%	5.8% 7.1%
Beta R-CNN [13]	GPU	ResNet-50	10.6%	47.1%	10.3%	6.4%
NOH-NMS [44]	GPU	ResNet-50	10.8%	53.0%	11.2%	6.6%
SKGNet (ours)	TeslaV100 GPU	ResNet-101	10.0%	**45.9%**	9.3%	**5.5%**

Current detection methods, including ours, mostly focus on improving detection accuracy while ignoring detection speed, which is not conducive to real-time detection. Besides, the real-world pedestrian conditions are usually very complex, so pedestrian detection in complex scenes remains challenging. As a result, future work should concentrate on increasing

pedestrian detection speed as well as improving pedestrian detection performance in complex scenes.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China (No. 2021ZD0112400), National Natural Science Foundation of China under Grant 61906032, the NSFC-Liaoning Province United Foundation under Grant U1908214, the Fundamental Research Funds for the Central Universities under grant DUT21TD107, the LiaoNing Revitalization Talents Program, No. XLYC2008017, and the Liaoning Key Research and Development Program under Grant 2019JH2/10100030.

REFERENCES

[1] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5187-5196.

[2] Y. Pang, J. Xie, M. H. Khan, et al., "Mask-guided attention network for occluded pedestrian detection," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4967-4975.

[3] J. Zhang, L. Lin, Y. Chen, et al., "CSID: center, scale, identity and density-aware pedestrian detection in a crowd," CoRR, abs/1910.09188, 2019.

[4] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," IEEE Trans. Pattern Anal. Mach. Intell, vol. 40, no. 4, pp. 973-986, Apr. 2018.

[5] J. Ren, X. Chen, J. Liu, et al., "Accurate single stage detector using recurrent rolling convolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5420-5428.

[6] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," European conference on computer vision. Springer, Cham, 2016, pp. 354-370.

[7] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3127-3136.

[8] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1904-1912.

[9] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3486-3495.

[10] Y. Pang, J. Cao, and X. Li, "Cascade learning by optimally partitioning," IEEE transactions on cybernetics, 2016, pp. 4148-4161.

[11] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, Jun. 2018, pp. 6995-7003.

[12] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 135-151.

[13] Z. Xu, B. Li, Y. Yuan, and A. Dang, "Beta r-cnn: Looking into pedestrian detection from another perspective," Advances in Neural Information Processing Systems, 2020.

[14] X. Wang, T. Xiao, Y. Jiang, et al., "Repulsion loss: Detecting pedestrians in a crowd," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7774-7783.

[15] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" European conference on computer vision, Springer, Cham, 2016, pp. 443-457.

[16] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE transactions on pattern analysis and machine intelligence, 36(8):1532-1545, 2014.

[17] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in Proceedings. international conference on image processing, IEEE, 2002, vol. 1, pp. I-I.

[18] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6459-6468.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, 2015. pp. 91-99.

[20] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1091-1100.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp. 580-587.

[22] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.

[23] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263-7271.

[24] W. Liu, D. Anguelov, D. Erhan, et al., "Ssd: Single shot multibox detector," European conference on computer vision. Springer, Cham, 2016, pp. 21-37.

[25] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," arXiv preprint arXiv:1509.04874, 2015.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[27] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "LFFD: A light and fast face detector for edge devices," 2019, arXiv:1904.10633.

[28] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 734-750.

[29] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 850-859.

[30] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint arXiv:1904.07850, 2019.

[31] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," arXiv preprint arXiv:1904.03797, 2019.

[32] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for singleshot object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 840-849.

[33] X.Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794-7803.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[35] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2015, pp. 2017-2025.

[36] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 510-519.

[37] X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," arXiv preprint arXiv: 1905.09646, 2019.

[38] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3213-3221.

[39] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," TPAMI, 34(4):743–761, April 2012.

[40] J. Deng, W. Dong, R. Socher, et al., "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248-255.

[41] W. Wang, "Adapted Center and Scale Prediction: More Stable and More Accurate," arXiv preprint arXiv: 2002.09053, 2020.

[42] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware r-cnn: detecting pedestrians in a crowd," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 637–653.

[43] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 618–634.

[44] P. Zhou, C. Zhou, P. Peng, et al., "NOH-NMS: Improving Pedestrian Detection by Nearby Objects Hallucination," in Proceedings of the 28th ACM International Conference on Multimedia. 2020, pp. 1967-1975.