Spatial Pyramid Attention for Deep Convolutional Neural Networks

Xu Ma, Jingda Guo, Andrew Sansom, Mara McGuire, Andrew Kalaani, Qi Chen, Sihai Tang, Qing Yang, *Senior Member, IEEE,* Song Fu, *Senior Member, IEEE*

***Abstract*—Attention mechanisms have shown great success in computer vision. However, the commonly used global average pooling in some implementations aggregates a three-dimensional feature map to a one-dimensional attention map, leading a significant loss of structural information in the attention learning. In this article, we present a novel Spatial Pyramid Attention Network (SPANet), which exploits the structural information and channel relationships for better feature representation. SPANet enhances a base network by adding Spatial Pyramid Attention (SPA) blocks laterally. By rethinking the self-attention mechanism design, we further present three topology structures of attention path connection for our SPANet. They can be flexibly applied to various CNN architectures. SPANet is conceptually simple but practically powerful. It uses both structural regularization and structural information to achieve better learning capability. We have comprehensively evaluated the performance of SPANet on four benchmark datasets for different visual tasks. The experimental results show that SPANet significantly improves the recognition accuracy without adding much computation overhead. Using SPANet, we achieve an improvement of 1.6% top-1 classification accuracy on the ImageNet 2012 benchmark based on ResNet50, and SPANet outperforms SENet and other attention methods. SPANet also significantly improves the object detection performance by a clear margin with negligible additional computation overhead. When applying SPANet to RetinaNet based on the ResNet50 backbone, we improve the performance of the baseline model by 2.3 mAP and the enhanced model outperforms SENet and GCNet by 1.1 mAP and 1.7 mAP respectively. The code of SPANet is made publicly available[1].**

***Index Terms*—Convolutional neural network, Attention mechanism, Spatial pyramid structure, Structural regularization, structural information, Image classification, Object detection.**

I. Introduction

In the last few years, we have witnessed a flourish of convolutional neural networks (CNNs) in computer vision research and applications. To improve the performance of CNNs, recent works add more convolutional layers to the CNN architectures. For instance, from 8-layer AlexNet [1] to 1000-layer ResNet [2], [3], in order to achieve higher accuracy for image recognition. However, more learnable layers introduce more parameters and prolong inference time.

Besides making the neural network deeper, there are recent efforts on developing attention mechanisms [4] for CNNs. An attention mechanism informs a CNN model where to look and what to pay attention to, making attention networks achieve better performance using fewer layers. SENet [5], for example, introduces Squeeze-and-Excitation (SE) blocks to consider channel dependencies in CNN. Non-local operations [6] describe the global context in a feature map.

Attention-based CNNs, such as SENet, CBAM, and RANet [5], [7], [8], employ the global average pooling (GAP) on feature maps. However, GAP aggregates a three-dimensional feature map into a one-dimensional attention map, which causes a loss of structural information in the intermediate feature maps. Moreover, applying GAP to every feature map emphasizes the effect of global structural regularization, but overlooks the detailed feature representation and structural information, especially when a feature map is large. For example, when we only apply the SE block to the first stage of ResNet (with a resolution of 56×56), the classification accuracy decreases, which indicates that an over-strong structural regularization is introduced while the detailed structural information is ignored. To mitigate the problem, the convolutional block attention module (CBAM) [7] explores channel-wise attention and spatial attention, which considers channel dependencies and structural information independently. However, simultaneously dealing with channel dependencies and structural information using multiple fully-connected layers and convolutional layers inevitably introduces more computation and latency. Furthermore, merely employing a large kernel convolutional layer for structural information extraction is inefficient and cannot capture long-range mutual correlation.

The issues caused by GAP make the shallow layers which output large-size feature maps unable to fully exploit the benefits from channel attention mechanisms [8]. To efficiently capture the structural information and channel relationship in attention modules, we creatively incorporate structural information in channel-wise attention blocks. In this article, we present a Spatial Pyramid Attention Network (SPANet), which introduces a spatial pyramid structure to encode the intermediate features and two point-wise convolutional layers to extract channel relationships. Our proposed SPANet has two major components: 1) a spatial pyramid structure which aggregates the feature contexts of three scales to combine the fine, coarse and global feature regularization, and 2) a combination

Xu Ma is with the the Department of Computer Science and Engineering, University of North Texas, Denton, TX, 76203 USA. He is also with the College of Information Science and Technology, Nanjing Forestry University, Nanjing, Jiangsu, China. E-mail: XuMa@my.unt.edu

J. Guo, Q. Chen, S. Tang, Q. Yang, and S. Fu are with the Department of Computer Science and Engineering, University of North Texas, Denton, TX, 76203 USA. E-mail: (JingdaGuo@my.unt.edu, QiChen@my.unt.edu, SihaiTang@my.unt.edu, qing.yang@unt.edu, song.fu@unt.edu).

A. Sansom is with the Department of Mathematics, University of North Texas, Denton, TX, 76203 USA. E-mail: (Andrew Sansom@my.unt.edu).

M. McGuire is with the Department of Mathematics and Statistics, Texas A&M University - Corpus Christi, Corpus Christi, TX 78412 USA. e-mail: (maracm17@gmail.com).

A. Kalaani is with the Department of Electrical and Computer Engineering, Georgia Southern University, Statesboro, GA 30458 USA. E-mail: (ak04526@georgiasouthern.edu).

[1]https://github.com/13952522076/SPANet_TMM

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068576, IEEE Transactions on Multimedia

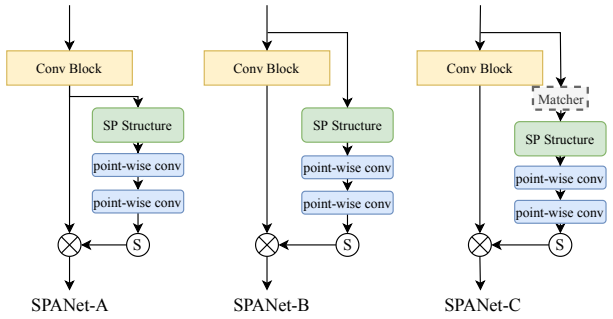


Fig. 1. The architecture of the spatial pyramid attention network (SPANet). We explore three variants of SPANet. SPANet-A learns attention from the current feature maps. SPANet-B learns from the previous feature maps. SPANet-C contains an optional channel matcher in the attention path.

of two point-wise convolutional layers and a sigmoid-based activation layer which encode and decode attention weights. Both components are designed to be lightweight.

Our spatial pyramid structure can be considered to bear certain similarities with SPPNet [9] and Region of Interesting Pooling [10] with regard to the pooling schema. The difference is that the spatial pyramid structure encodes a feature map with more structural information, while SPPNet and Region of Interesting Pooling produce a fixed-length feature vector. In addition to the capability of retaining spatial information in each channel, an intriguing property of our spatial pyramid structure is that it only introduces three parameters to reweigh the values of the scales which are then aggregated by summation. We call this operation weighted summation. The plugged SPA module in each block adaptively adjusts the importance of three scales of structural information. All structural information in the spatial pyramid structure is not learnable (achieved by pooling operations), making them almost cost-free. This small computation overhead contributes to SPANet's improved performance.

Inspired by the connection design of self-attention and recent works on cross-layer operations [11], [12], we present three topology structures of the Spatial Pyramid Attention module in our proposed SPANet, referred to as SPANet-A, SPANet-B, and SPANet-C. SPANet-A learns attention from current feature maps, which follows a traditional self-attention path connection schema. SPANet-B learns from previous feature maps by a cross-layer pattern, which is independent of the current features. SPANet-C adds an optional channel number matcher to SPANet-B for channel number matching. Figure 1 depicts the three topology structures of SPANet.

We comprehensively evaluate the performance of SPANet on four benchmark datasets, i.e., CIFAR-100 [13], down-sampled ImageNet [14], ImageNet [15], and MS COCO [16]. Without bells and whistles, SPANet outperforms the state-of-the-art works [2], [5], [17]–[19]. Experimental results show that the structural information in the attention mechanism, which this article focuses on, is crucial for models' performance. Applying our SPANet to the base ResNet50, we achieves a 1.6% accuracy improvement on the ImageNet benchmark. Similar improvement is also observed on the other classification datasets, showing that SPANet's effectiveness is not confined to one particular dataset. Moreover, SPANet is general and applicable to other vision tasks in addition to image classification. When applied to object detection, SPANet consistently achieves a significant mean Average Precision (mAP) improvement over baseline models. More specifically, we improve the detection performance by 2.3% mAP over ResNet50 using the RetinaNet detector and improve the performance of Cascaded R-CNN by 2.2% mAP, with only negligible additional computation overhead.

The rest of the article is organized as follows. Section II discusses the related work. Section III describes the design of SPANet, including the SPA module, attention path connection, and Spatial Pyramid Attention. The performance of SPANet for image classification and object detection is evaluated in Section IV. Section V concludes the article with remarks on future research.

II. Related Work

Deep Neural Networks. Convolutional neural networks has shown promising results for many vision tasks, including image recognition [2], [9], object detection [20], [21] and instance segmentation [22]. In these years, various works have demonstrated the benefits of increasing network depth, from AlexNet [1] to DenseNet [17]. Most recently, some tentative efforts have increased the network depth to over 10,000 layers [23]. Rather than focusing on the accuracy improvements obtained from deeper architecture, some efforts have been made towards lightweight networks that yield comparable accuracy using fewer parameters and FLOPs [18], [24]–[27]. For example, to investigate the balance among depth, width, and resolution, EfficientNet [28] employs a compound coefficient to scale network width, depth, and resolution uniformly. It achieves state-of-the-art performances with minimal cost. In this paper, we are in pursuit of a lightweight module that can be seamlessly integrated with various deep or lightweight models to improve networks' representational ability further.

Multi-Path Connection. Multi-path connection scheme was first used in Highway Networks [29], [30], if not earlier. By allowing unimpeded information flowed across several layers, a Highway Network can reuse the information from previous layers, which facilities the training of deep networks. Moreover, gating units are employed to regulate information flow. Subsequently, He *et al.* proposed Residual Networks (ResNet) [2], [3], which learn the residual functions by adding skip-connections. The ResNet shows that an identity mapping shortcut is crucial to ease the optimization [2], [3]. Hence, ResNet discards the gating units used in Highway Networks and keeps the information passed through shortcuts. The promising performance achieved by ResNet has made shortcut connections attractive. As a more dense reformulation, DenseNet [17] connects every convolutional layer in a deep convolutional network. Without introducing more parameters, it effectively alleviates the vanishing gradient problem and improves feature reuse.

In addition to shortcut connections, there are works studying the internal multi-path connections in convolutional blocks [31]. The InceptionV4 Network [31] is one of this kind.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068576, IEEE Transactions on Multimedia

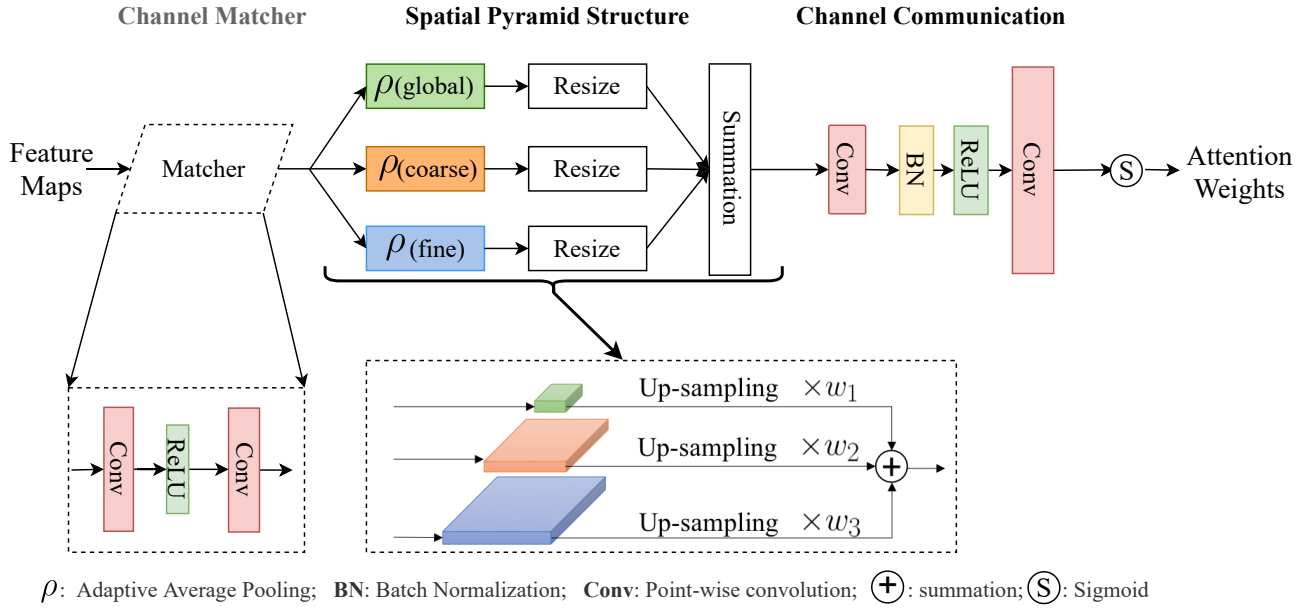


Fig. 2. The architecture of our spatial pyramid attention module. It consists of three major components, i.e., channel matcher, spatial pyramid structure, and channel communication. Channel matcher is mainly designed for SPANet-C to match the channel number and integrate channel information. The spatial pyramid structure includes adaptive average pooling of three different sizes to integrate structural regularization and structural information in an attention path. The channel communication component learns an attention map from the output of the spatial pyramid structure. In SPANet-A, the input feature map is the current output of a block. In SPANet-B and SPANet-C, the input feature map is the previous output of a convolutional block.

Besides a shortcut connection, each inception block in InceptionV4 contains 3-6 carefully designed paths. All these paths are integrated using filter concatenation as input to the next block. More recently, attention-based networks such as SENet [5] and CBAM [7] provide an independent attention path to learn the weight of each channel and achieve state-of-the-art performance.

Attention Mechanisms. Attention mechanisms [4] have been prevailed in computer vision for years [32]. By adopting a gating function (such as softmax and sigmoid) or adding complementary information, attention mechanism can selectively emphasize salient features and suppress insignificant features. Thus, visual features could be better captured and exploited. In [5], a Squeeze-and-Extraction block was proposed to learn the channel-wise attention for each convolutional layer, which provides an end-to-end training paradigm for the channel attention learning. Inspired by SENet, Competitive-SENet [33] studies attention from both the residual path and the shortcut path. Although Competitive-SENet achieves promising performance, it is tailored particularly for Residual Networks [2], limiting its generalization to other models. Without being limited to channel attention, Sanghyun Woo *et al.* [7] exploited the relation between channel-wise attention and spatial attention and proposed a Convolutional Block Attention Module (CBAM). CBAM is composed of two parts, i.e., a channel-wise attention block and a spatial attention block. The two attention blocks in CBAM can tell what (channel) to look and where (spatial) to focus on. Unlike CBAM that learns channel-wise attention and spatial attention separately, our proposed SPANet learns channel-wise and spatial attention in an integrated fashion and requires much fewer computations and parameters.

Multi-Scale Feature Fusion. Multi-scale feature fusion holds prevalence for a long time. One of the most typical examples is the spatial pyramid pooling method [9]. To address the limitation of fixed input size of earlier CNNs, He et al. [9] proposed a spatial pyramid pooling strategy that adaptively pools the feature maps to three scales and feeds them to the finial fully-connected layers. The simple but effective strategy dramatically improves the CNN capability for image classification and object detection. Similar to [9], Feature Pyramid Network [34], Cascaded R-CNN [20] and Res2Net [35] also consider the use of multi-scale intermediate feature maps for low-level/high-level feature semantics fusion and feature map resolution compensation. Most recently, HRNet [36] provides a new scheme for multi-scale feature fusion by concatenating feature maps of different resolutions. In this article, we learn from these works and leverage the multi-scale aggregation for our local and global context fusion.

III. Spatial Pyramid Attention (SPANet)

In this section, we present the design details of spatial pyramid attention (SPANet), a plug-in module that is compatible with various base CNN networks. SPANet models the global context and local context of different scales using a pyramid structure. Then two point-wise convolutional layers are employed to explore the channel relationships. The attention map is then up-sampled to the sample size of the original feature map. We adopt the dot-product self-attention formulation to refine the original feature map. Fig. 2 depicts the design of our method.

The SPA module can be applied after each convolutional layer. For efficiency, we add the SPA module after each block in a backbone network. As an example, we add 8

SPA modules in ResNet18 and 16 modules in ResNet50. We keep the pyramid scales the same for all the modules. Such a dynamic pipeline of scale combinations can lead to a better performance. Meanwhile, the pipeline structure also requires more hand-crafted design (or network architecture searching), which makes the network more complex. As a result, we investigate the effectiveness of our SPA module. In the following sections, we elaborate the design of each component in SPANet.

A. Design of SPA Module

We investigate three different scales to explore different extent of structural regularization. They are then integrated by a weighted summation. We then use two point-wise convolutional layers to explore the channel relationship for each pixel. The generated attention map is then up-sampled and considered as the activation of the original feature map. Our design is generic in that the SPA module can be applied to a variety of base models.

Assume a CNN architecture has L layers and each layer generates a feature map x. The output of the l-th layer could be denoted by x_l. The index of a layer, l, is within $[1, L]$. We use $\rho(\cdot)$ to represent an adaptive average pooling layer, $\tau(\cdot)$ is for a point-wise convolutional layer, and $\sigma(\cdot)$ is a sigmoid activation function.

For an intermediate feature map $x_l \in \mathbb{R}^{C\times W\times H}$, a CNN model with an attention mechanism takes x_l as input, learns an attention map, and generates the output by combining the attention map and the original feature map. The output generated by our Spatial Pyramid Structure component can be expressed as follows.

$$\mathbf{S} = w^{\mathsf{T}} \left[\rho_{fine}(x_l), \rho_{coarse}(x_l), \rho_{global}(x_l)\right], \tag{1}$$

where $w \in \mathbb{R}^3$ is a learnable parameter vector used for the weighted summation.

Inspired by [9], we design 3-level pyramid features: fine, coarse, and global, which is achieved by average pooling operations. Hence, no learnable parameter is introduced in feature extraction. Remarkably, the three scales represent three different structural regularization terms.

For ease of our discussion and a focus on the design of SPANet, we take away the batch normalization and activation layer for now. The transformation $\mathbf{T}$ made by the SPA module can be described as follows.

$$\mathbf{T} = \mathbf{U}\left(\sigma\left(\tau\left(\tau\left(\mathbf{S}\right)\right)\right)\right), \tag{2}$$

where function $\mathbf{U}(\cdot)$ is employed to upsample the attention map to match with the shape of the original feature map. The output of SPANet $\mathbf{T}$ is then combined with the original feature map by element-wise multiplication.

Note that Equation (2) only showcases the fundamental transformation conducted by the SPA module. The other components, e.g., the batch normalization layer and activation layers, are included in our implementation whose performance is evaluated in Section IV.

B. Attention Connection Scheme

A typical design pattern of self-attention based networks is that an attention map is learned from the current feature map and then applied to the feature map itself [5], [37]. That design pattern, however, limits the exploration of attention path connections. To tackle this problem, we investigate the topology of attention path connections and then design three spatial pyramid structures for SPANet, i.e., SPANet-A, SPANet-B, and SPANet-C, which are depicted in Figure 1. Next, we will describe these three structures in detail.

In **SPANet-A**, the current feature map x_l goes through the attention path, from which an attention map is created. The transformation performed by SPANet-A is as follows.

$$x_l = \mathbf{T}\left(x_l\right) \otimes x_l, \tag{3}$$

where $\otimes$ represents element-wise multiplication. From Equation (3), we can see the self-attention path connection in SPANet-A is similar to the traditional schema.

In **SPANet-B**, an attention map is learned from x_{l-1}, where $x_{l-1} \in \mathbb{R}^{C'\times W'\times H'}$ is the input of x_l in **SPANet-A**. The output of SPANet-B can be described as

$$x = \mathbf{T}\left(x_{l-1}\right) \otimes x_l. \tag{4}$$

SPANet-B enables the attention path to learn more generalized weights, as the attention path and the original convolutional block path are independent of each other. On the other hand, the two paths are not entirely irrelevant. They are trained jointly.

When the channel number in x_{l-1} is different from the channel number in x_l, the attention path might not produce the most accurate weights for x_l. So, an optional channel matcher is added to the beginning of the attention path in **SPANet-C**, if $C' \neq C$. The channel matcher consists of two point-wise convolutional layers, a batch normalization layer, and the ReLU function. Thus, the output of SPANet-C can be expressed as

$$x = \mathbf{T}\left(\psi\left(x_{l-1}\right)\right) \otimes x_l, \tag{5}$$

where $\psi(\cdot)$ represents the channel matcher component. The purpose of adding the matcher is to integrate channel information and match channel numbers of the output feature maps. It makes the attention path further independent of x_l.

The three variants of SPANet can be applied to a variety of CNN architectures. In this section, we present the topology structure of attention path connections. Next, we will describe how these attention mechanisms are implemented.

C. Major Components of SPANet

Global average pooling (GAP) has been used to aggregate an input feature map into a one-dimensional vector in many attention networks [5], [7], [7], [33]. While the application of GAP achieves structural regularization [38], the detailed structural information is missing. SPANet aims to achieve structural regularization and explore structural information at the same time. To this end, we develop the spatial pyramid structure, which performs the average pooling with three different sizes. Figure 2 presents the architecture of the spatial pyramid attention module.

1) Spatial Pyramid Structure: Global average pooling (GAP) aggregates the global information in each channel. It has been widely used in computer vision: image recognition [2], object detection [39], semantic segmentation [40], etc. Attention-based CNNs, such as [5], [7], [8], use GAP on each feature map. In terms of the functionality, GAP is like a structural regularizer and can prevent overfitting [38]. On the other hand, applying GAP to every feature map overemphasizes the effect of regularization and overlooks the feature representation and structural information, especially when a feature map is large. For instance, if a 56×56 feature map is aggregated into a mean value, we lose features' representation capability, affecting the performance of feature learning.

To tackle this problem, we design a spatial pyramid structure that is used in attention blocks. Our proposed spatial pyramid structure adaptively pools an input feature map to three different scales: A fine average pooling is to capture more feature representation and structural information. A coarse average pooling aims at a trade-off between the structural information and structural regularization. A 1×1 average pooling is the traditional GAP providing a strong structural regularization. The three outputs are then upsampled to a spatial resolution as the fine representation, and combined using a weighted summation. Such a small learnable combination makes the network adaptively adjusts the importance of these three constituents in each attention block. With our designed spatial pyramid structure, we aggregate the local context and the global context in an integrated manner and improve the feature representation ability significantly.

2) Channel Relation: The spatial pyramid structure produces an attention map (denoted by $\mathbf{S}$), which is a combination of the outputs from three pooling layers. However, it could not directly be employed for the exploration of channel dependency.

To solve these problems, we explore the channel relationship [5], which encodes S and produces a transformed attention map. Specifically, the transformation block has two point-wise convolutional layers, and a sigmoid function is employed to re-scale the output into a range of $(0, 1)$. We set the intermediate output channel number as C/r, where r is the reduction rate with a default value of 16.

Note that the resolution of the attention map obtained here is not identical to that of the original feature map. To match the resolution, we upsample the attention map to the same resolution as the original feature map. Hence, SPANet could be considered as a regional-aware attention design.

3) Channel Matcher: The attention path in SPANet-B (Section III-B) learns from the input of the convolutional block with a channel number of C'. However, it is used to activate the feature map with a channel number of C. The mismatch of the number of channels may lead to a discrepancy in attention learning, affecting the performance of SPANet.

We solve this problem in SPANet-C (Section III-B) where a channel matcher is used when the input channel number is not equal to the output channel number. Particularly, our channel matcher consists of two point-wise convolutional layers, which are lightweight. The intermediate channel number is set to $\max(C', C)/r$.

D. Efficiency

All the learnable layers introduced in our SPA module are point-wise convolutional layers (operated on small feature maps) and the Batch-Normalization layer. Besides, a learnable vector $w \in \mathbb{R}^3$ is employed to aggregate the feature pyramid. Hence, SPANet only introduces a small amount of extra computation overhead and parameters to the base model. When applying SPA to ResNet50, we only add 0.14G Flops and 2.51M parameters. This additional overhead is negligible compared with the Flops and parameters in ResNet50 (4.122G and 25.557M respectively).

In the next section, we will show and discuss the performance of SPANet from our comprehensive experimentation on four benchmark datasets and two vision tasks, i.e., image recognition and object detection.

IV. Experimental Results and Evaluation

In our experiments, SPANet is applied to several state-of-the-art convolutional neural networks. We compare its performance with that of the existing attention modules, such as SENet [5], GENet [19] and CBAM [7]. We use CIFAR-100 [13], Downsampled ImageNet [14] and ImageNet [41] benchmark datasets in image recognition experiments. The performance of object detection using SPANet is also studied on the MS COCO [16] dataset. To achieve a deep understanding of SPANet, we have conducted extensive ablation studies as well.

A. Datasets

We use four widely used datasets in our experiments: CIFAR-100[2], Downsampled ImageNet[3], ImageNet 2012[4], and MS COCO[5], which are publicly available.

CIFAR-100 includes 60,000 colored images (32×32 pixels) belonging to 100 classes. Each class contains 600 images. We use 500 images in training, and 100 images in testing.

Downsampled ImageNet is a downsampled version of the original ImageNet dataset. It contains all the images in ImageNet and they are re-sized to 32×32 for computational efficiency. Overall, it has 1,281,167 training images and 50,000 validation images in 1,000 classes.

ImageNet dataset is a large-scale labeled dataset organized according to the WordNet architecture. It includes 1,281,167 training images, and 50,000 verification pictures, belonging to 1000 categories. The number of images in each class varies and the resolution of the images is not the same. In recent years, ImageNet serves as a benchmark dataset for the image recognition task. In our experiments, we adopt the dataset published in 2012.

MS COCO (Microsoft Common Objects in Context) originated from the Microsoft COCO dataset that Microsoft funded and annotated in 2014. Similar to the ImageNet competition, it is regarded as one of the most prevalent and authoritative

[2]https://www.cs.toronto.edu/ kriz/cifar.html
[3]http://image-net.org/small/download.php
[4]http://www.image-net.org/challenges/LSVRC/2012/downloads
[5]https://cocodataset.org

TABLE I
CLASSIFICATION ACCURACY ON CIFAR-100.

Module	MobileNetV2	DenseNet	VGG16
Base	74.97	74.72	72.99
SE	75.38	74.67	72.61
GE	75.93	73.63	73.01
SPA-A	75.73	74.71	72.97
SPA-B	**76.11**	75.43	72.88
SPA-C	75.85	**75.87**	**73.14**

TABLE II
CLASSIFICATION ACCURACY ON THE DOWNSAMPLED IMAGENET.

Module	MobileNetv2		ResNet18	
	top-1 acc.	top-5 acc.	top-1 acc.	top-5 acc.
Base	51.41	76.35	53.25	76.98
SE	51.60	76.31	53.54	77.32
SPA-A	**52.06**	76.62	**53.64**	**77.62**
SPA-B	51.77	76.63	53.30	77.32
SPA-C	52.02	**76.67**	53.47	77.39

competitions in the computer vision field. The COCO data set is large and prosperous object detection, segmentation, and captioning data set. This data set aims at scene understanding, which is mainly intercepted from complex daily scenes. There are 80 categories and more than 330,000 images, of which 200,000 are labeled. The number of individuals in the entire dataset exceeds 1.5 million. We employ the MS COCO dataset mainly to evaluate the performance of SPANet in object detection.

B. Classification Performance on CIFAR-100

We have implemented SPANet using the PyTorch [42] framework and evaluated its performance. A stochastic gradient descent method is employed to train all models in the experiments. We apply a 0.9 Nesterov momentum and a $5e^{-4}$ weight decay. In our experiments, we use the CIFAR-100 dataset with a batch size of 256. The initial learning rate is set to 0.1, and the learning rate is decreased by a factor of 10 for every 70 epochs. The epoch size is 300. We exploit a data augmentation approach used in [2], [3] in the training. Specifically, an original image is padded with four pixels of zeros on each side. After that, it is randomly cropped to 32×32 pixels. Then, half of the generated images are horizontally flipped in random. To facilitate model training, we normalize the image data by using channels' means and standard deviations.

In our experiments, we compare the performance of SPANet with that of SENet and GENet [19]. Three different base networks are studied, including VGG16 [43], DenseNet [17], and the light-weight model MobileNetV2 [18]. For the spatial pyramid structure in SPANet, we employ the resolution of 4×4, 2×2, and 1×1 for the fine, coarse, and global representation respectively since the original image's resolution is low. The image recognition results on CIFAR-100 are presented in Table I.

From the table, we observe that SPANet achieves the best recognition accuracy in most (not all) scenarios. We improve the performance of MobileNetV2 by 1.14% (SPA-B) and enhance the performance of DenseNet by 1.15%. Compared to the SE and GE counterparts, SPA outperforms them by a clear margin.

However, SPANet does not achieve the best results on VGG16. The results of SPA, SE, and GE do not display major statistical difference from that of the baseline VGG16. As we discuss before, GAP improves the structural regularization and tackles over-fitting. The proposed spatial pyramid structure leverages both structural regularization and structural information for enhanced learning capability. This may also lead to over-fitting on smaller datasets. Our experimental results show the training loss from SPANet approaches zero on CIFAR-100. Thus its performance becomes stable. This result infers that a larger training set leads to better performance. When applying SPANet and SENet to VGG16, we improve the performance of vanilla VGG16 on the downsampled ImageNet by 0.77% and 0.41%, top-1 accuracy reaching 53.04% and 52.71% respectively.

C. Classification performance on Downsampled ImageNet

SPANet is also evaluated on a down-sampled ImageNet dataset (resolution: 32×32). We take the same data processing scheme as that on the CIFAR-100 dataset. To evaluate the generality of SPANet, we employ different base networks for the downsampled ImageNet dataset: MobileNetv2 and ResNet18. We show the top-1 and top-5 classification accuracy in Table II.

From the table, we have the following findings.

- SPANet outperforms all the base models and the SE counterparts. Moreover, all three types of SPANet achieve higher accuracy than the base models, and 4 out of the 6 SPA models outperform SENet. These results indicate that SPANet is more effective. Therefore, using both structural regularization and structural information can significantly improve the performance of attention mechanisms.
- The best performance is not achieved by one particular type of SPANet in all the cases. For instance, SPANet-A achieves the highest top-1 accuracy for MobileNetv2 and ResNet18, while SPANet-C has the best top-5 accuracy for MobileNetv2. Furthermore, all these three designs achieve better results and possess different performance gains. This indicates the necessity of improving the topology structure of attention path connections.

D. Experimental Results on ImageNet

We further evaluate the performance of SPANet on the full-scale ImageNet [15] benchmark. Following the common practice [2], we perform random-size cropping of images to 224×224, and horizontally randomly flip images with a probability of 0.5. We train networks from scratch using synchronous SGD with a weight decay of 0.0001 and a momentum of 0.9. All experiments are conducted on a server with 8 Tesla V100 GPUs. For ResNet [2] and its variants, we

TABLE III
SINGLE CROP CLASSIFICATION ACCURACY (%) ON THE IMAGENET VALIDATION SET.

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)	FLOPs (G)	Parameters (M)
ResNet50 [2]	75.8974	92.7224	4.122	25.557
SE-ResNet50 [5]	77.2877	93.6478	4.130	28.088
GE-ResNet50 [19]	76.2357	92.9847	4.127	25.557
CBAM-ResNet50 [7]	77.2840	93.6005	4.139	28.090
SPA-ResNet50(ours)	**77.4880**	**93.6098**	4.262	28.074
MobileNetV2 [2]	71.0320	90.0670	0.320	3.505
SE-MobileNetV2 [5]	72.0482	90.5812	0.321	3.563
GE-MobileNetV2 [19]	72.2776	90.9120	0.322	3.555
CBAM-MobileNetV2 [7]	71.9069	90.5114	0.324	3.565
SPA-MobileNetV2(ours)	**72.5438**	**91.1344**	0.325	3.562

train the networks for 100 epochs with a batch size of 256 (i.e., 32 images per GPU), stating at a learning rate of 0.1 and decreasing it by ten every 30 epochs. For lightweight CNN models such as MobileNetV2 [18], we train the networks for 150 epochs with a batch size of 512 (i.e., 64 images per GPU), stating at a learning rate of 0.1 and adjusting it using a cosine decay method [44].

We select two base CNN models: ResNet50 and MobileNetV2, as representatives of a normal model and a lightweight model. For comparison, we select several state-of-the-art self-attention modules, including SENet [5], GENet [19], and CBAM [7]. The metrics that we measure are Top-1 and Top-5 accuracy for evaluating the performance of the aforementioned modules, and the numbers of FLOPs and parameters for comparing the efficiency of the tested modules. For convenience, we use the SPA-A variant in our experiments and denote it as SPA. Considering the large resolution of the ImageNet dataset, we employ the three scales of 7×7, 4×4, and 1×1 in the spatial pyramid structure.

As shown in Table III, our SPANet outperforms the baseline models by a large margin and consistently dominates the top performance on ImageNet compared with other attention methods. By applying our SPANet, we improve the performance of ResNet50 by 1.6% and that of MobileNet by 1.5%. Remarkably, our SPANet is lightweight, and only introduces 2.5M and 0.06M parameters, and negligible Flops to ResNet50 and MobileNetv2, respectively. Such a small additional computation overhead from our SPA module is justified by the significant improvement of model's performance.

E. Ablation Studies

To better understand the inherent properties of our SPANet, we have conducted comprehensive ablation studies. We extensively study each component and provide a deep understanding of the internal operations of SPANet.

1) Attention Connection: Unlike the traditional self-attention mechanisms that learn from the input feature map and activate the feature map, SPANet explores two more attention connection schemes in SPANet-B and SPANet-C, respectively. To analyze the effects from the cross-layer attention connection, we apply SPANet-B and SPANet-C to SENet, denoted as SE+ and SE++. We employ ResNet as our base model and carry out experiments on the CIFAR-100 dataset. The experimental results are presented in Table IV.

TABLE IV
IMPACT OF CONNECTION SCHEMES AND NETWORK DEPTH USING THE CIFAR-100 DATASET.

Network	18 layers	50 layers	101 layers
SE	75.34	79.4	**79.52**
SE+	75.24	79.18	79.23
SE++	**75.41**	**79.45**	79.44
SPA-A	**75.96**	**80.32**	79.3
SPA-B	75.68	79.95	79.12
SPA-C	75.33	79.77	79.33

As shown in Table IV, different connection schemes present different properties. Experimental results show that SE++ outperforms SE on both ResNet18 and ResNet50. However, SE surpasses SE++ when the network goes to 101 layers. In addition, we find that SE++ consistently achieves better results than SE+. However, we does not observe a similar phenomenon on SPANet, and SPA-B slightly outperforms SPANet-C. As an example, SPA-B-ResNet18 outperforms SPA-C-ResNet18 by 0.35% (75.68% *vs.* 75.33%). These findings indicate that the topology structure of an attention path connection should not be confined to only one schema, and exploration of the topology structure helps achieve better performance.

2) Training and Testing Loss: We also present the training loss and the testing loss of SPANet on different base networks over multiple datasets.

Figure 3 plots the training loss and validation loss on the ImageNet dataset. Clearly, our SPANet decreases the training loss and validation loss of base ResNet50 by a clear margin. Compared to other state-of-the-art methods like SENet, GENet, and CBAM, our SPANet consistently yields the lowest training/validation loss. Moreover, the loss of our SPANet is much more stable during the training.

The effects of our SPANet are not just akin to a particular dataset. We also present the training loss of MobileNetv2 and SPANet counterparts on the CIFAR-100 dataset in Figure 4. Intuitively, even on the tiny dataset, CIFAR-100, all of the three SPANet variants consistently yield lower loss than vanilla MobileNetv2. Similar phenomena are also shown on the down-sampled ImageNet dataset and other base models, suggesting that SPANet is able to provide a better feature representation for many CNN architectures.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068576, IEEE Transactions on Multimedia

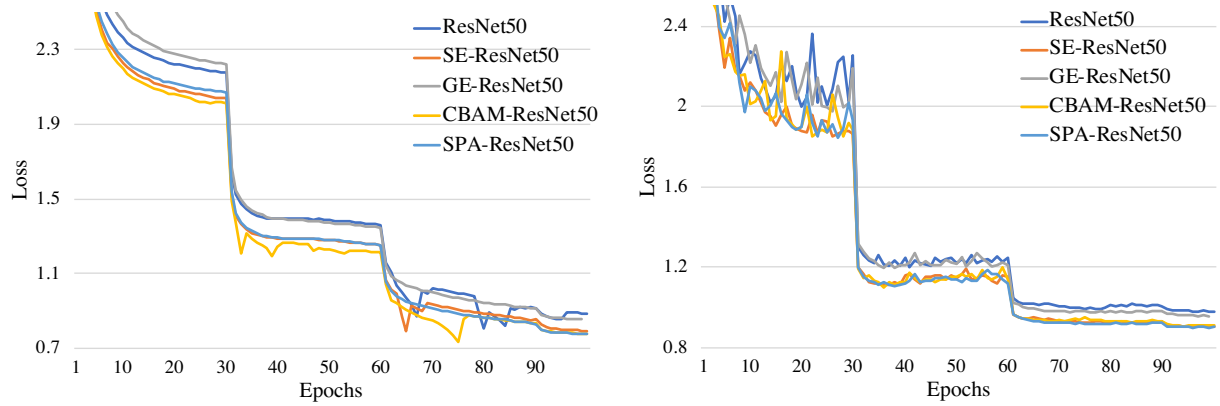


Fig. 3. Training loss (left) and validation loss (right) on the ImageNet dataset. Compared with other state-of-the-art methods, SPANet consistently achieves the minimal loss on both the training set and validation set.

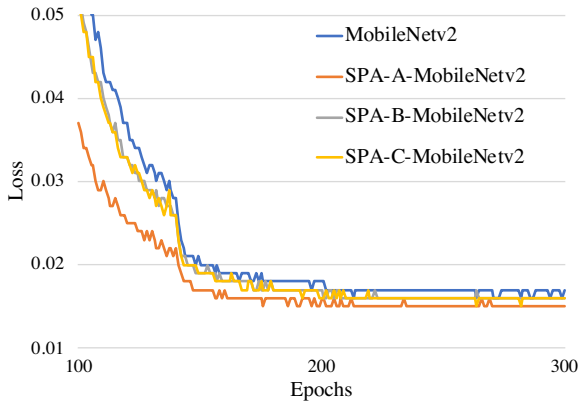


Fig. 4. Training loss of MobileNetv2 and the corresponding variants on the CIFAR-100 dataset. For better visualization, we present the loss starting at the 100th epoch.

TABLE V
IMPACT OF THE SCALE USING THE IMAGENET DATASET.

Scales	top-1 accuracy	top-5 accuracy
baseline	75.90	92.72
[1, 2]	77.29	93.52
[1, 2, 4]	77.33	93.60
[1, 4, 7]	**77.49**	93.61
[1, 2, 4, 7]	77.43	**93.62**

3) Spatial Pyramid Structure: The presented spatial pyramid structure provides a substitute for GAP. In this ablation study, we evaluate the performance of SPANet-A under different combinations of three feature scales. Table V presents the numerical results.

As shown in Table V, the influence of different scale combinations is not significant, and they all surpass the baseline by a clear margin. Specifically, the combination of 1×1, 4×4 and 7×7 achieves the best result (77.49 top-1 classification accuracy). When more scales, e.g., [1, 2, 4, 7], are used, the performance improvement is not significant.

4) Heatmap Visualization: We visually investigate the SPA module. We present the heatmap of the last convolutional layers in a ResNet50 model and the corresponding variants using Grad-CAM [45]. We compare our SPANet module with SE-ResNet50 and the vanilla ResNet50. The results are plotted in Fig. 5.

The experimental results show that all models correctly pay attention to the target objects and ignore the trivial regions on all the three input images. The models exhibit different characteristics. We can see that vanilla ResNet50 pays attention to a big region that encompasses the target object, while SE-ResNet50 pays attention to the center part of an object. Different from them, our SPA-ResNet50 accurately focuses on the most discriminative parts of an object. SPANet biases the location of the most informative and discriminative features and simultaneously suppresses the irrelevant regions. These features and results of SPANet meet our design expectation.

From all these ablation studies, we can see SPANet consistently outperforms the baseline and SENet based networks. These results verify that it is imperative to combine the structural information and structural regularization with attention paths in convolutional neural networks in order to achieve accurate image recognition (as discussed in Section III-C).

F. Object Detection Performance on MS COCO

We have comprehensively evaluated the effectiveness of our SPANet for image recognition in the preceding sections. Please note that our SPANet is not confined to only one particular task. Instead, it can serve as a plug-in module for various types of computer vision tasks. Here, we evaluate the performance of SPANet in object detection.

In this set of experiments, we use the MS COCO dataset. We employ two state-of-the-art detection architectures, i.e., Cascade R-CNN [20] and RetinaNet [46]. All object detection experiments are built on the open-source mmdetection framework [48]. We also employ the Feature Pyramid Network [34] to obtain a richer representation by extracting features from different layers in the backbone. We measure the average precision (AP) of bounding box detection under different conditions on the challenging COCO validation dataset [16]. All input images are re-scaled to ensure the shorter side has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068576, IEEE Transactions on Multimedia

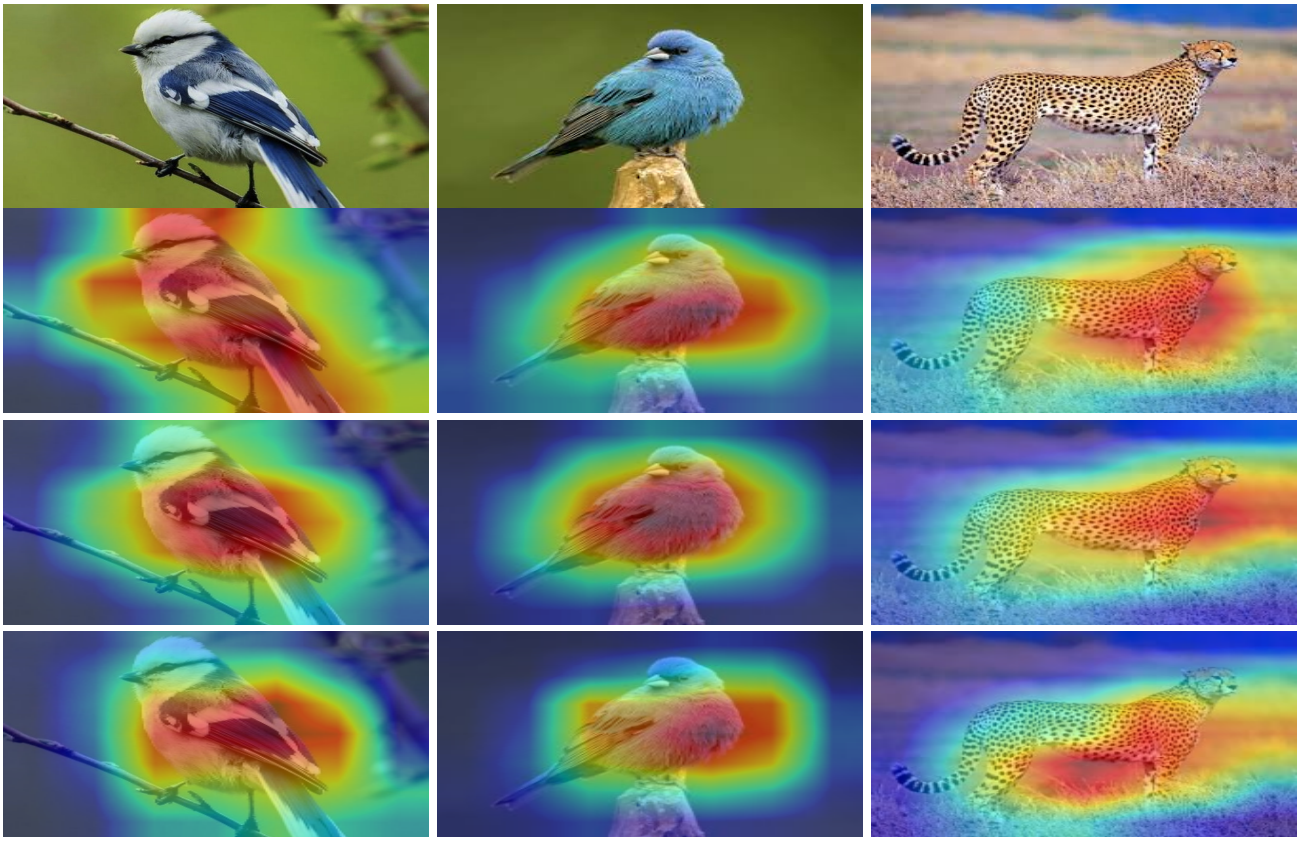


Fig. 5. Heatmap visualization results. **Top row:** original images; **Second row:** heatmap of vanilla ResNet50 (The vanilla ResNet50 pays attention to a big region that encompasses the target object.); **Third row:** heatmap of SE-ResNet50 (SE-ResNet50 pays attention to the center part of an object.); **Bottom row:** heatmap of our SPA-ResNet50 (SPA-ResNet50 presents different attentions and focuses on the most discriminative parts of an object). Best viewed in color.

TABLE VI
OBJECT DETECTION PERFORMANCE (%) WITH DIFFERENT BACKBONES ON THE MS-COCO VALIDATION DATASET. THE BEST PERFORMANCE IS HIGHLIGHTED IN "**BOLD**".

Detector	Backbone	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	GMac	Parameters (M)
RetinaNet [46]	ResNet50 [2]	36.2	55.9	38.5	19.4	39.8	48.3	239.32	37.74
	SE-ResNet50 [5]	37.4	57.8	39.8	20.6	40.8	50.3	239.43	40.25
	SPA-ResNet50 (ours)	**38.5**	**59.2**	**41.0**	**22.1**	**42.3**	**50.7**	239.54	40.25
Cascade R-CNN [20]	ResNet50 [2]	40.6	58.9	44.2	22.4	43.7	54.7	234.71	69.17
	GC-ResNet50 [47]	41.1	59.7	44.6	23.6	44.1	54.3	234.82	71.69
	SPA-ResNet50 (ours)	**42.8**	**61.6**	**46.3**	**24.6**	**46.3**	**56.9**	234.93	71.67

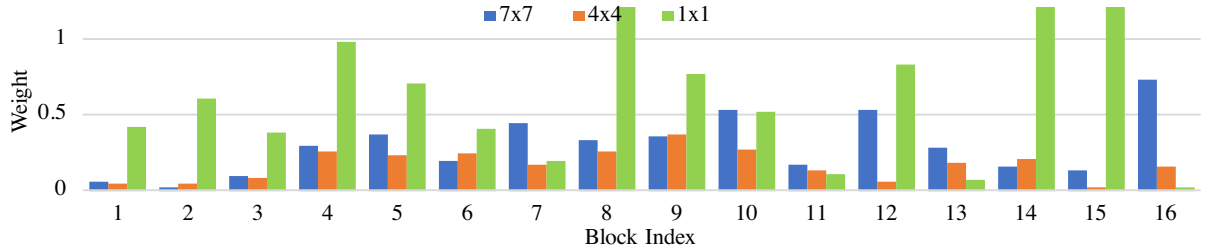


Fig. 6. The absolute value of weight w in our spatial pyramid structure. We find that all three scales contribute to the spatial pyramid structure in our SPA module.

800 pixels. We adopt the similar settings used in Faster R-CNN [21] and train all models with a total of 16 images per batch (i.e., two images per GPU). For the backbone model, we select ResNet50 and its variants. Note that all the backbones are pre-trained on the ImageNet benchmark, using those listed in Table III.

We train all the detection models for 24 epochs using the synchronized SGD optimizer with a weight decay of 0.0001 and a momentum of 0.9. The learning rate is set to 0.02 for Cascade R-CNN and 0.01 for RetinaNet as used in [20], [46], and decreased by a factor of 10 at the 18th and 22nd epochs during the training. For clarity, We employ the SPANet-A counterpart for comparison.

Table VI presents the object detection results. From the table, we can find that SPANet improves the detection performance of the baseline models by a clear margin for both RetinaNet and Cascaded R-CNN. More specifically, SPANet improves the average precision of RetinaNet by 2.3 mAP and Cascaded R-CNN by 2.2 mAP. Surprisingly, SPANet outperforms both SENet and GCNet significantly, i.e., 38.5 mAP *vs.* 37.4 mAP and 42.8 mAP *vs.* 41.1 mAP. Four object detection examples are shown in Figure 7. Moreover, SPANet consistently outperforms other methods in terms of all evaluated metrics.

This encouraging improvement are attributed to the regional-aware design in the spatial pyramid structure. By dividing the feature map into $n \times n$ individual patches, where each patch considers the local (fine), neighbor (coarse) and global contexts simultaneously, SPANet intrinsically performs a regional-aware activation for a detection backbone network. To evaluate this aspect, we explicitly depict the absolute value of the weights for the three scales in each block. In Figure 6, we find that all of the three different scales of feature representations contribute to the spatial pyramid structure in our SPANet to generate a better result. Moreover, the global representation contributes the most in all blocks except for the last block, and the weight of fine representation keeps increasing when the index of convolutional block increases. This property leads us to rethink how to better bridge the gap between the CNN backbone and detector head in a detection framework in our future research.

In Fig. 6, we find that the weight of the 1×1 scale in the last block is very small, while the weights of the other two finer scales in the first two blocks are small. This is caused by the dynamic feature aggregation. Specifically, after receiving the input feature maps, SPANet aggregates contextual data at multiple scales, including local (fine), neighbor (coarse) and global contexts. In the first two stages, as the feature representations are relatively shallow, aggregation mainly focuses on the local (fine) context. The global context (e.g., the 1×1 pooling representation) aggregated in later stages can help solve the problem. In the last block, the resolution of feature maps is greatly reduced and contexts at all of the three scales contain sufficient global information. The 1×1 pooling representation is suppressed in the last block.

V. Conclusions

In this article, we present the Spatial Pyramid Attention Network (SPANet), a novel design to enhance the performance of CNNs. SPANet incorporates the spatial pyramid structure which integrates the structural information and structural regularization, and further explores the channel relationship. Moreover, we investigate the topology structure of attention path connections and develop three types of SPANet with different connection schemes. Experimental results on four datasets (i.e., CIFAR-100, Downsampled ImageNet, ImageNet 2012, and MS COCO) show the efficiency and effectiveness of SPANet for image classification and other vision tasks (We present the results from object detection). The significant performance improvement inspires us to investigate the relationship between feature pyramid networks (for object detection) and SPANet. In the current SPANet, the design of spatial pyramid structure (i.e., fine, coarse and global representations) is simple. How to design a more general and dynamic spatial pyramid structure is an interesting problem. As a future work, we will develop dynamic spatial pyramid structures and further improve SPANet for better performance by looking into the regional-aware activation.

Acknowledgment

The work has been supported in part by the National Science Foundation grants CNS-1852134, OAC-2017564, ECCS-2010332, and the research grant from Fujitsu Laboratories of America Inc. A preliminary version of this article was accepted by 2020 IEEE International Conference on Multimedia and Expo (ICME) and received the Best Student Paper Award [49].

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[6] 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*, 2018, pp. 7794–7803.

[7] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[8] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

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

[11] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," in *Advances in Neural Information Processing Systems*, 2019, pp. 1307–1318.

[12] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11030–11039.

[13] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068576, IEEE Transactions on Multimedia

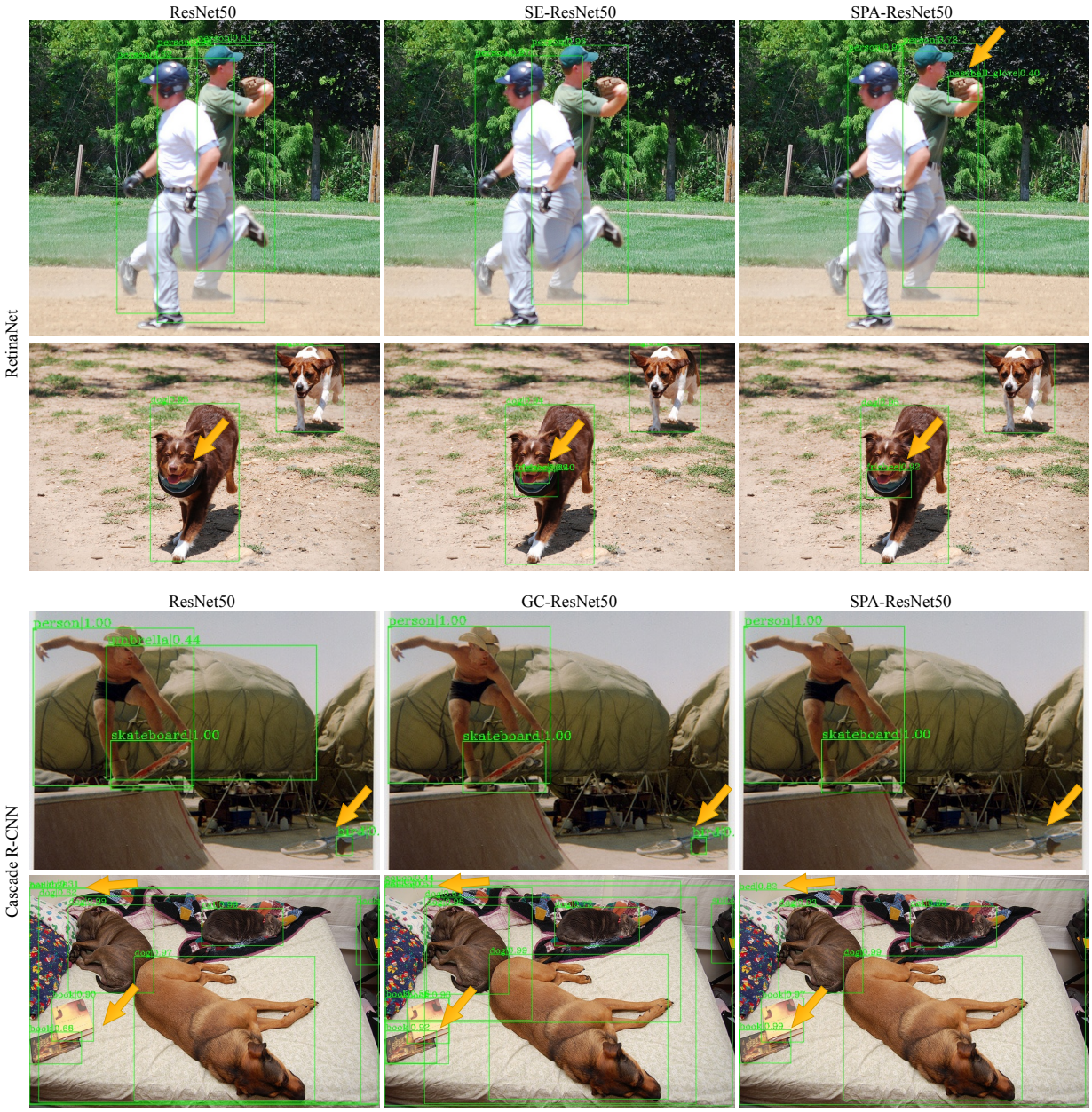


Fig. 7. Examples of the object detection results. **Top 2 rows are the results from RetinaNet with different backbones:** from left to right are ResNet50, SE-ResNet50 and SPA-ResNet50; **Bottom 2 rows are the results from Cascade R-CNN with different backbones:** from left to right are ResNet50, GC-ResNet50 and SPA-ResNet50. We highlight the most discriminative parts using yellow arrows. The results show SPANet achieves better detection performance than the rest.

[14] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *arXiv preprint arXiv:1707.08819*, 2017.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[19] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Advances in neural information processing systems*, 2018, pp. 9401–9411.

[20] Zhaowei Cai and Nuno Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer*

vision, 2017, pp. 2961–2969.

[23] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W Cottrell, and Julian McAuley, "Rezero is all you need: Fast convergence at large depth," *arXiv preprint arXiv:2003.04887*, 2020.

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[25] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[26] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[27] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828.

[28] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[29] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[30] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

[31] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[32] Tam V Nguyen, Qi Zhao, and Shuicheng Yan, "Attentive systems: A survey," *International Journal of Computer Vision*, vol. 126, no. 1, pp. 86–110, 2018.

[33] Yang Hu, Guihua Wen, Mingnan Luo, Dan Dai, Jiajiong Ma, and Zhiwen Yu, "Competitive inner-imaging squeeze and excitation for residual network," *arXiv preprint arXiv:1807.08920*, 2018.

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[35] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[36] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[37] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3286–3295.

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[39] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[40] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[43] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[44] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

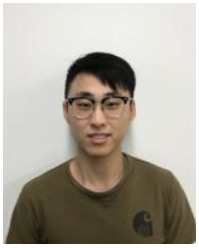
[46] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[47] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[48] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al., "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[49] Jingda Guo, Xu Ma, Andrew Sansom, Mara McGuire, Andrew Kalaani, Qi Chen, Sihai Tang, Qing Yang, and Song Fu, "Spanet: Spatial pyramid attention network for enhanced image recognition," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

Xu Ma is a Ph.D candidate in the Department of Computer Science and Engineering at the University of North Texas, Denton, Texas. He received his B.S. degree and M.S. degree in College of Information Science and Technology at Nanjing Forestry University, China, in 2015 and 2018, respectively. His research interests include deep learning, computer vision, pattern recognition, and connected autonomous vehicles.

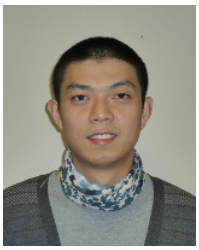


Jingda Guo is a Ph.D. candidate in the Department of Computer Science and Engineering at the University of North Texas, Denton, Texas. He received his B.S. degree in Electrical Engineering from Northeast Electric Power University, China, and M.S. degree in Computer Engineering from University of Delaware, USA, in 2015 and 2017, respectively. His research interests include Internet of Things, connected autonomous vehicles, and computer vision.

Andrew Sansom is an undergraduate student in Mathematics and Computer Science at the University of North Texas, Denton, Texas. He is interested in artificial intelligence and has participated in a research project to enhance convolutional neural networks for autonomous vehicles. He will graduate in Spring 2021.

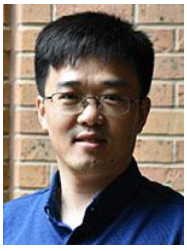
Mara McGuire is currently a Master's student in the Department of Mathematics and Statistics at Texas A&M University - Corpus Christi, Corpus Christi, Texas. She received her B.S. degree in Applied Mathematics from Texas A&M University - Corpus Christi in 2020. She is interested in the mathematical foundation of deep learning and participated in a project on vehicular edge computing and security at the University of North Texas.

Andrew Kalaani is currently an undergraduate student in the Department of Electrical and Computer Engineering at Georgia Southern University, Statesboro, Georgia. His research interests include deep learning, autonomous vehicles, communication systems, and circuits. He participated in a project to optimizing classification accuracy using CNNs for autonomous vehicles at the University of North Texas.

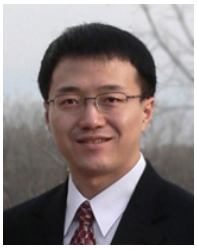


Qi Chen received B.S. and M.S. degrees in Electronic Engineering from Xidian University and Northwestern Polytechnical University, China, in 2006 and 2010, respectively. He received his Ph.D. degree in Computer Science and Engineering at the University of North Texas, Denton, Texas, in 2020. His research interests include perception on autonomous vehicles, edge AI and Internet of Things.

Sihai Tang is a Ph.D. candidate and a part of the Dependable Computing Systems Laboratory and the Connected and Autonomous Vehicles Laboratory in the Department of Computer Science and Engineering at University of North Texas, Denton, Texas. He received his B.S. degree in Computer Science from the University of Texas at Austin in 2017, Austin, TX, USA. His research interests include edge computing, federated machine learning, and storage systems.



Qing Yang is an Assistant Professor in the Department of Computer Science and Engineering at the University of North Texas, Denton, Texas. He directs the Connected and Autonomous Vehicles Laboratory. He received B.S. and M.S. degrees in Computer Science from Nankai University and Harbin Institute of Technology, China, in 2003 and 2005, respectively. He received his Ph.D. degree in Computer Science from Auburn University in 2011. His research interests include Internet of Things, connected and autonomous vehicles, network security and privacy. He serves as an Associate Editor for IEEE Internet of Things, and Elsevier Vehicular Communications journals. He is a Senior Member of IEEE.



Song Fu is an Associate Professor in the Department of Computer Science and Engineering at the University of North Texas, Denton, Texas. He directs the Dependable Computing Systems Laboratory and co-directs the Connected and Autonomous Vehicles Laboratory. He received B.S. and M.S. degrees in Computer Science from Nanjing University of Aeronautics and Astronautics and Nanjing University, China, in 1999 and 2002, respectively. He received his Ph.D. degree in Computer Engineering from Wayne State University in 2008. His research interests include parallel and distributed systems, machine learning, cloud and edge computing, connected and autonomous vehicles, system reliability, and high performance computing. He has published over 120 journal and conference articles in these areas. He is a Senior Member of IEEE, and Member of ACM.