Engineering Applications of Artificial Intelligence 148 (2025) 110456

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Residual-like multi-kernel block and dynamic attention for deep neural networks

Hanxiang Wang [a,b,c], Yanfen Li [a,*], Tan N. Nguyen [d], L. Minh Dang [e,f,**]

[a] School of Computer Science, Qufu Normal University, Rizhao, China
[b] Shandong Provincial Key Laboratory of Data Security and Intelligent Computing, China
[c] Rizhao-Qufu Normal University Joint Technology Transfer Center, China
[d] Department of Architectural Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea
[e] Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam
[f] Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural network
Attention
Multi-scale
Residual block

ABSTRACT

Traditional network architectures struggled with a uniform approach to receptive field (RF) sizes, leading to suboptimal performance across scales. Although recent advances have addressed the problem by utilizing different RF sizes, a balance between accuracy and complexity remains elusive. In addition, the existing group attention mechanism that simply uses the squeeze-and-excitation method neglects the spatial position information in the feature selection and fusion process. Therefore, this research introduces a lightweight and efficient architecture named Split-Dense Adaptive Network (SDANet) to cope with these limitations. In the proposed network, a residual-like multi-kernel method is implemented to enable better feature extraction under diverse RF sizes. Next, a new grouped attention module processes features dynamically and highlight the location information. Also, the constructed feature augmentation structure strengthens the model's representation. Furthermore, a new channel split and merge strategy is utilized for computation reduction. Compared with state-of-the-art methods, our model achieved better generalization ability, less computational complexity, and superior precision based on various public datasets. The introduced network shows a promising general applicability in the field of computer vision, and further inspires research on supervised deep learning.

1. Introduction

The representation of image features from the initial pixel representation to the later feature descriptors (scale invariant feature transform (SIFT) (Lowe, 2004), speeded-up robust features (SURF) (Bay et al., 2008), etc.) is to explore the most effective information expression method. In the last few decades, deep learning-based convolutional neural network (CNN) approaches have demonstrated excellence in conquering diverse difficulties for various cognition tasks such as fine-grained classification, object detection, and semantic segmentation, which require special attention to the contextual information of multi-scale target patterns. Thus, the design of a deep learning model with multi-scale feature representation is critical to processing different objects in natural scenes.

Considering the fine-to-coarse mode of input features, multi-scale representations have been realized by numerous backbone networks in order to improve performance. In earlier research, VGGNet (Simonyan and Zisserman, 2014) increases the network depth and expands receptive fields (RFs) via a simple stack of convolutional operations. Even though VGGNet with a deep structure can extract features on a large scale, it causes a relatively fixed RF in each layer. Later, some consequent structures, such as the residual modules (Xie et al., 2017; He et al., 2016), use short connections, making more excellent layer-wise representations. In addition, the multi-kernel approach was applied to improve model's recognition performance (de Lima et al., 2016; Atif et al., 2023), but this method results in huge computational complexity. Different from most approaches that strengthen layer-wise multi-scale processing ability, an innovative method called Res2Net is put forward to process multi-scale features at a more granular level (Gao et al., 2019a). Motivated by the success of the state-of-the-art (SOTA)

* Corresponding author.
** Corresponding author. Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam.
E-mail addresses: yanfen@qfnu.edu.cn (Y. Li), danglienminh@duytan.edu.vn (L.M. Dang).

https://doi.org/10.1016/j.engappai.2025.110456
Received 13 March 2024; Received in revised form 25 December 2024; Accepted 26 February 2025
Available online 4 March 2025

networks, this study aims to improve precision and efficiency by optimizing the original backbone structure. Specifically, a residual-like multi-kernel module with a dynamic feature augmentation structure is introduced to deal with the multi-scale feature extraction and representation problem. Compared with prior work, the proposed module can perceive features under various RFs without overfitting and computation overhead.

Recently, the attention mechanism brought considerable benefits at low computational costs for an expansive range of visual tasks (Zhang et al., 2018a; Xu et al., 2018; Minh et al., 2022). Since the attention mechanism can emphasize meaningful features and suppress less important parts through a reasonable weight allocation, many efforts have been made to incorporate attention modules into CNNs for performance gains. In context with current research trends, most attention modules are investigated from two basic dimensions: channel and spatial dimensions. A representative example of channel attention is squeeze-and-excitation network (SENet) (Hu et al., 2018), in which a gating system is constructed to recalibrate features adaptively. Similarly, spatial attention is introduced with channel attention in Bottleneck attention module (BAM) (Park et al., 2018) and Convolutional block attention module (CBAM) (Woo et al., 2018). Apart from channel and spatial attention, some dynamic approaches (Chen et al., 2020; Li et al., 2019) are applied to obtain features in multiple branches. Nevertheless, the existing feature selection process ignores the positional information and produces more parameters for the whole network. Therefore, we propose a new attention module in this research to mitigate the problems caused by extra computations and inefficient feature representations. Our constructed attention differs from the aforementioned attention modules as it utilizes one-dimensional pooling operation from two spatial directions. This method efficiently obtains position information in multiple branches with fewer computational resources.

In this research, we make an effort to design a novel architecture that is able to realize an adequate representation performance without sacrificing computation costs. The main contributions are analyzed from the following four aspects.

- Propose a residual-like multi-kernel approach for better feature extraction.
- Design a grouped attention mechanism to select and fuse features dynamically.
- Present a feature augmentation structure to strengthen the model's representation ability.
- Introduce an effective channel split and merge strategy with an adjustable coefficient to reduce computational costs.

The rest of this article is arranged as follows. A critical review of relevant studies is discussed in Section 2. The effectual architecture used for this research is introduced in Section3. Section 4 describes the details of the experimental dataset. In addition, extensive experiments for image classification, object detection, and segmentation are presented to demonstrate our contributions. Section 5 comprises conclusions and future research lines.

2. Related work

Deep architectures with residual block. Significant advancements in the existing deep neural networks have extensively inspire the research on visual pattern recognition. In the last decade, quite a few architectures have been introduced for the purpose of elucidating the close correlation between network depth and learning ability (Simonyan and Zisserman, 2014; Szegedy et al., 2015). Yet, the traditional models that simply stack blocks or modules with the same topology gain inferior training results when the network depths increase. As for this problem, a residual structure was presented to support the training of a deeper neural network (He et al., 2016). Compared to previous plain nets, residual nets (ResNets) are more easily optimized, and their training performance is considerably boosted with the increase of the depth. One year later, the module called ResNeXt, which adopts the multi-branch strategy in residual blocks, was constructed to expand the capacity of the original ResNet (Xie et al., 2017). Based on ResNet, another block named Res2Net was presented by creating multiple connections in a single residual module (Gao et al., 2019a). The multi-scale operation of Res2net is irrelevant to other existing layer-wise approaches for representation enhancement. Experiments proved that Res2Net has wider RFs and a more powerful multi-scale representation. Beyond these SOTA residual structures, this research designs a residual-like multi-kernel block to refine features under different RFs.

Attention-guided networks. The attention mechanism is a momentous module in neural networks, whether learning salient features or processing negligible signals. Its superiority has been demonstrated in recent studies regarding various attention-guided models. For example, Squeeze-and-Excitation (SE) block is produced for channel-wise feature recalibration, which can be flexibly and directly applied to standard convolution networks or other complicated transformations (Hu et al., 2018). Consequently, BAM (Park et al., 2018) and CBAM (Woo et al., 2018) emerged as channel and spatial attention modules. In addition, a dynamic attention mechanism that adaptively chooses RF sizes of neurons was designed and integrated into CNNs to form an attention-guided Selective Kernel Network (SKNet) (Li et al., 2019). Most of these studies focus on utilizing two-dimensional global pooling to encode channel or spatial signals, which pay less attention to the target position in the spatial axis. In contrast, this paper proposes a simple grouped attention module that can extract adequate and accurate positional information from a multi-branch.

Grouped/Depthwise/Dilated convolutions. By introducing a new dimension (cardinality), group convolutions can reduce the computational complexity of regular convolutions and even their training error (Xie et al., 2017; Krizhevsky et al., 2012). ConvNeXt demonstrated strong performance in ImageNet classification tasks, but its effectiveness is not significant in the segmentation task (Liu et al., 2022). And the use of a large number of regular convolutions in its grouped structure results in high computational costs and model complexity. As a particular case of grouped convolution, depthwise separable convolutions with the same groups and channels are presented in Xception (Carreira et al., 1998) and MobileNetV1 (Howard et al., 2017) in order to disintegrate normal convolutions into depthwise and pointwise convolutions. Pyramid Vision Transformer (PVT) utilized non-overlapping patch sequences and depthwise convolutions, which can disrupt the local continuity of the image and affect the model's ability to capture local features (Wang et al., 2021a). Also, PVT uses fixed length position encoding and cannot flexibly process multi-scale images. Moreover, dilated convolutions take action in the spatial direction instead of channel-wise expansion of grouped convolutions (Li et al., 2019; Liu and Moon, 2021). Apart from the above three approaches to strengthen the model's learning ability with fewer parameters, Wang et al. propose a Cross Stage Partial Network (CSPNet) that can alleviate enormous inference costs without learning ability reduction by incorporating features throughout the whole network (Wang et al., 2020). Rather than the strided convolution operation used in CSPNet for downsampling, we adopt a softpooling function to retain details. In addition, a novel fusion form is introduced in this study to augment the feature-sharing capability of CSPNet.

3. Methodology

The proposed Split-Dense Adaptive Network (SDANet) is a flexible, practical, and efficient architecture that can be seamlessly plugged into any network to describe multi-scale patterns for diverse visual tasks such as recognition, detection, and segmentation. In this section, an original residual-like multi-kernel learning approach for feature extraction is first introduced in Section 3.1. Then, the design concept of the grouped attention module for adaptive feature selection is explained in detail

(Section 3.2). After that, we describe some algorithms that are dedicated to augmenting features in Section 3.3. Finally, a channel split and merge structure with an adjustable coefficient is analyzed to reduce the model's computations (See details in Section 3.4). The overall architecture of the proposed SDANet is shown in Fig. 1.

3.1. Feature extraction

A flexible feature extraction scheme that extracts valuable information from multiple scales is conducive to promoting the model's learning and representation. In contrast to the regular stacked architectures, residual structures can capture high-level semantic information of raw inputs in a stable manner, and it can avoid the vanishing gradient problem (Xie et al., 2017; He et al., 2016). Currently, Res2Net emerged as the most representative residual structure due to its excellent performance in different tasks. Accordingly, we construct a residual-like multi-kernel structure that adopts the Res2Net connection strategy to remain powerful multi-scale feature extraction ability.

Fig. 2 (a) illustrates the connection strategy of Res2Net, and the detailed structure of the proposed SDANet block is shown in Fig. 2(b). Although the Res2Net connection mode enhances the information flow in the feature extraction process, the model still requires robust contextual information abstraction for better results under complicated scenes (Gao et al., 2019a). Therefore, the traditional 3x3 convolutions of the Res2Net block are replaced by convolutions with different kernel sizes to extract different scales of feature maps under different RFs. In this regard, the large RF can process images with wide information distribution, while the small RF is suitable for images with more local information (Luo et al., 2016). Since dilated convolutions can expand the RF with less parameters without sacrificing resolutions of feature maps, we adjust the dilation rates of dilated convolutions to realize the function of standard convolutions with different kernel sizes. The actual kernel size of the dilation convolution is $(d-1)*(k-1)+k$, where d and k represent the dilation rate and kernel size, respectively. It should be noted that the dilation convolution is standard convolution when $d=1$. When d increases, the actual kernel of the dilation convolution becomes larger, and the RF of the corresponding convolution operation also becomes larger. In order to strike a balance between computational load and model performance, we adopted three dilation rates ($d=1$, 2, 3) corresponding to three convolution sizes (3x3, 5x5, 7x7).

The multi-scale feature fusion can promote the model to capture more global and local information in the whole information flow (Wang et al., 2021b, 2022). In this research, the dense structure with cross-branch feature fusions is introduced to fully use the deep features from different branches, as shown by the dotted arrow in Fig. 2 (b). The multi-scale feature fusion approach of SDANet is shown in Equation (1), where x_i is the input feature map of each branch, and the range interval of the branch index i is [1, b]. b represents the specific number of branches, and it is initially set to the same number (4) as in the Res2Net for this example, and its value can be adjusted to expand or reduce branches of the module. When b increases, the model's parameters and computational complexity also increase. K and O refer to the corresponding convolution operation and output in the branch, respectively.

$$O_i=\begin{cases} x_i & i=1; \\ K_i(x_i+O_{i-1}) & i=2; \\ K_i(x_i+O_{i-1}+O_{i-2}) & 3\le i\le b. \end{cases} \tag{1}$$

3.2. Feature selection

The fusion between multi-scale features easily brings about information dilution and aliasing effect. An attention mechanism that can establish channel relationships and analyze global information is significant in addressing the above problems (Luo et al., 2022). Therefore, this research introduces a lightweight attention mechanism for feature correction by calculating the attention in two different directions. As shown in Fig. 2 (b), the feature maps $[O_1, O_2, O_3, O_4] \in \mathbb{R}^{H\times W\times C/4}$ from four branches with different operations are input into the proposed grouped attention network (GANet). The recalibrated features $[\widehat{O}_1, \widehat{O}_2, \widehat{O}_3, \widehat{O}_4] \in \mathbb{R}^{H\times W\times C/4}$ are concatenated along the channel direction and then transferred from GANet to a 1x1 convolution layer.

$$U=(H\times W)^{-1}\sum_{x=1}^{H}\sum_{y=1}^{W}O(x,y) \tag{2}$$

A detailed process is shown in Fig. 2 (c) to illustrate how GANet selects and updates features. Four groups of feature maps are first

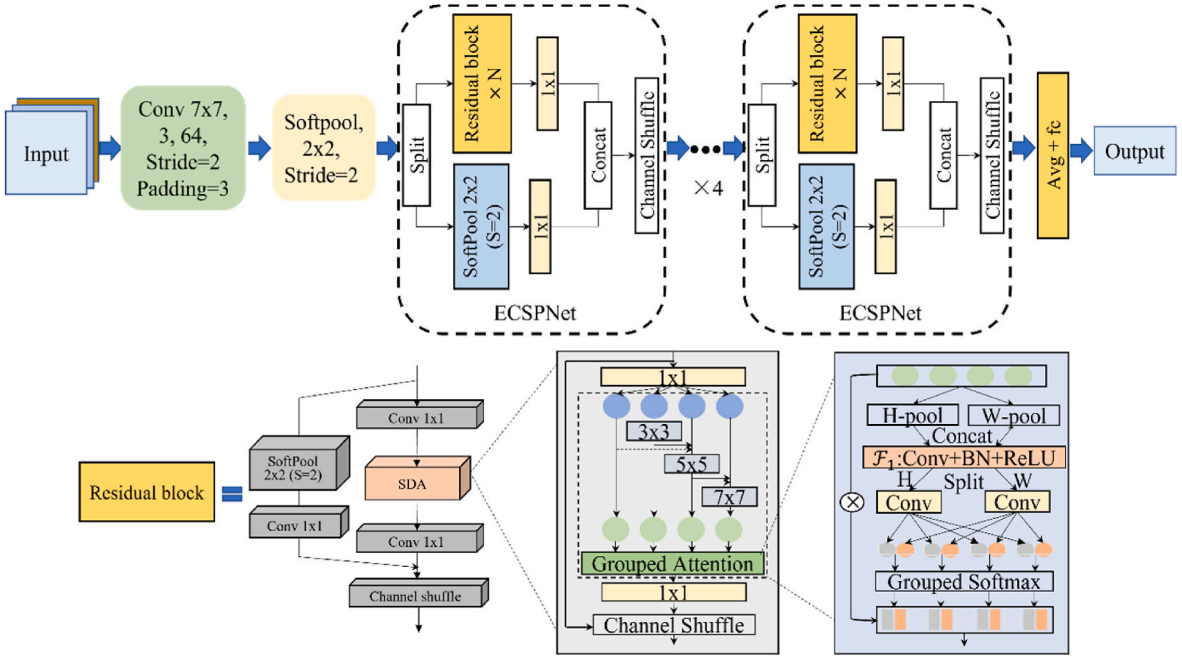

Fig. 1. The overall architecture of the presented SDANet, which includes the feature extraction (SDA), feature selection (Grouped Attention), feature augmentation (channel shuffle + softpool), and computation reduction (ECSPNet).

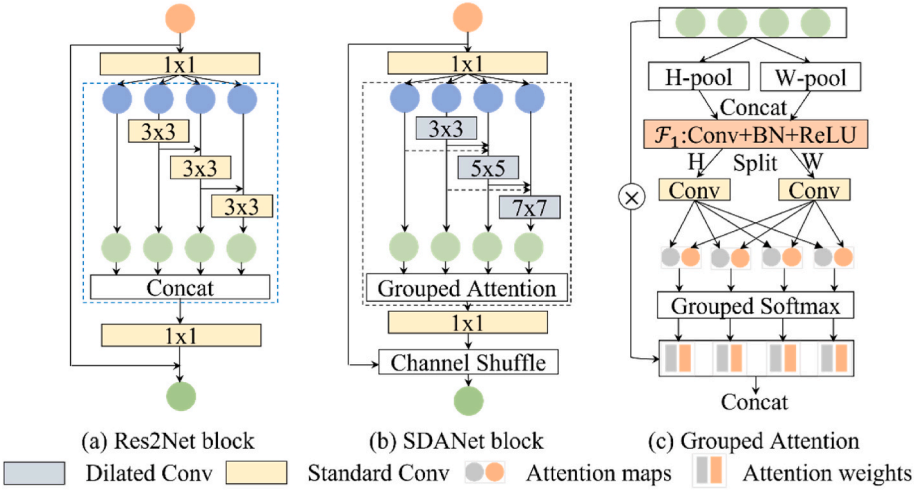

Fig. 2. The proposed residual block and grouped attention module. **Note:** In this study, the structure in the dotted blue box of Res2Net block (a) is called Res2, and the structure in the dotted black box of the SDANet block (b) is named SDA. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

combined by the addition operation, and then the grouped tensor is input to two one-dimensional (1D) generalized-mean pooling (GMP) layers. The GMP operations with kernel sizes of $H \times 1$ and $1 \times W$ can obtain the information from the horizontal and vertical directions, respectively. Unlike the traditional channel encoding methods (Hu et al., 2018; Li et al., 2019) that adopts 2D global pooling to squeeze the tensor O based on the spatial dimension $H \times W$ in Equation (2), we encode channels in two spatial axes to generate the direction-aware feature vectors as expressed in Equations (3) and (4). Hence, the proposed encoding approach can offer location information in the spatial dimension.

$$U_H = W^{-1} \sum_{x=1}^{W} O(H, x) \tag{3}$$

$$U_W = H^{-1} \sum_{y=1}^{H} O(y, W) \tag{4}$$

After that, the feature vectors $U_H \in \mathbb{R}^{H\times 1\times C/4}$ and $U_W \in \mathbb{R}^{w\times 1\times C/4}$ are concatenated into a new vector $U' \in \mathbb{R}^{(H+W)\times 1\times C/4}$. The generated vector is sent to a transformation $\mathcal{F}_1 : U' \rightarrow Z \in \mathbb{R}^{(H+W)\times C'}$ in order to learn inter-channel dependencies. C' is the number of output channels, and its calculation is shown in Equation (5), where b represents the base width of a convolution layer.

$$C' = \lfloor C/2 \times (b/64) \rfloor \tag{5}$$

The feature vector Z is separated into two independent vectors $Z_H \in \mathbb{R}^{H\times C'}$ and $Z_W \in \mathbb{R}^{W\times C'}$. Subsequently, these two vectors are sent to the 1×1 convolution layers to recover the channel numbers, and they are separated into four groups of attention maps corresponding to the inputs. Four attention maps $[Z_i^H, Z_i^W]$ are used to create four sets of attention weights $[A_i^H, A_i^W]$, including horizontal and vertical directions, via the softmax activation function in Equations (6) and (7). Different from the split attention used in ResNeSt (Zhang et al., 2022), which uses softmax to capture relationship the among feature maps within the same group, our grouped attention leverages the feature correlations between different branches to promote a better feature selection. Additionally, our grouped softmax computes channel weights from two spatial directions, which are then weighted to the input features. Here, i represents the index of the branch. $\omega \in \mathbb{R}^{B\times C/4}$ are the weights of B branches, and $B = 4$ for this example.

$$A_i^H = \frac{\exp\left(\omega Z_i^H\right)}{\sum_{j=1}^{B} \exp\left(\omega Z_j^H\right)} \tag{6}$$

$$A_i^W = \frac{\exp\left(\omega Z_i^W\right)}{\sum_{j=1}^{B} \exp\left(\omega Z_j^W\right)} \tag{7}$$

Finally, four sets of attention weights are used to transform the input $[O_1, O_2, O_3, O_4] \in \mathbb{R}^{H\times W\times C/4}$ to $[\widehat{O}_1, \widehat{O}_2, \widehat{O}_3, \widehat{O}_4] \in \mathbb{R}^{H\times W\times C/4}$ by:

$$\widehat{O}_i = O_i \times A_i^H \times A_i^W \tag{8}$$

3.3. Feature augmentation

In the feature extraction process, the deep learning model continuously applies the downsampling technique to feature maps to increase RFs and relieve the storage pressure (Zhou, 2020). There are two methods to realize downsampling, which include the pooling operation and strided convolution operation. In deep learning-based architectures, the models that use the strided convolution layer for downsampling can learn nonlinear features better than the models with pooling operation (Gao et al., 2019b). Therefore, the 1×1 convolution with the stride of 2 is widely used in ResNet and its related variant models to achieve the purpose of downsampling as shown in Fig. 3 (b). However, the convolution operation in which the stride is larger than the kernel size causes lots of details in feature maps to be ignored. The above problem can be alleviated by supplying a 2×2 pooling layer with the stride of 2 before the convolutional layer, whose stride is adjusted from 2 to 1 (He et al., 2019). Instead of maxpooling or avgpooling, this study adopts the softpooling that generates activation values for the kernel region using softmax as shown in Fig. 3 (d). Because the softpooling can retain more feature information while reducing computational overhead (Stergiou et al., 2021).

As discussed in Section 3.1, the proposed SDA adopts the multi-branch strategy achieved by the channel split. Yet the features of each branch become independent after the channel split, which blocks the fusion of the features from different branches (Krizhevsky and Hinton, 2009). For this issue, a channel shuffle operation is added after the feature concatenation to disrupt feature nodes and realize information integration, as shown in Fig. 3(c) and (d). In addition, the channel shuffle used in (Ma et al., 2018; Zhang et al., 2018b) is a random

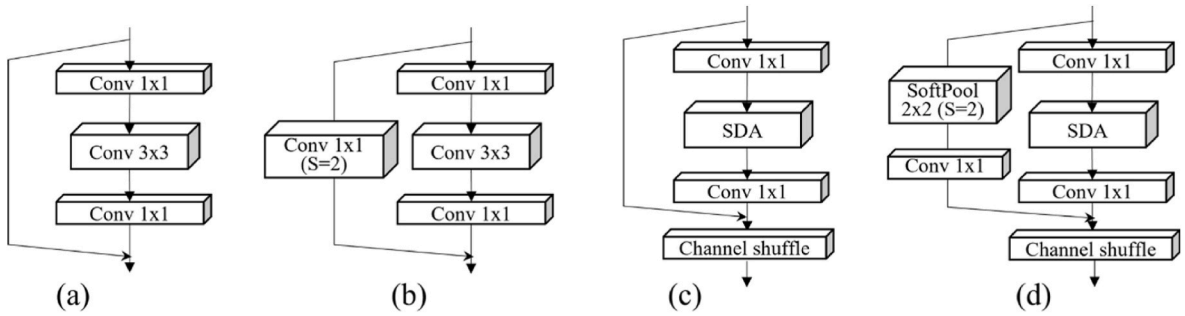

Fig. 3. Residual blocks of ResNet and the proposed SDANet. (a) The basic ResNet residual unit; (b) downsampling residual unit (2 ×) of ResNet; (c) the basic residual unit of the SDANet; (d) downsampling residual unit (2 ×) of the SDANet. Note that, S is stride.

operation, which reduces the correlation between channels and inevitably causes noises. To mitigate the negative effects noise, we improved the channel shuffle algorithm by adding the image processing technique (Gaussian filtering).

3.4. Computation reduction

To reduce the deep learning network's computational burden without performance decline, some approaches have been carried out in recent studies (Xie et al., 2017; Howard et al., 2017). In this context, the CSPNet is superior to concurrent work owing to two advantages: The model's parameters and inference computations can remarkably decrease by diminishing the repeated gradient information; Besides, the CSPNet considerably enhances the representation ability. In this research, an enhanced CSPNet (ECSPNet) is designed and integrated into the proposed SDANet to further improve both accuracy and time efficiency. As shown in Fig. 4 (b), the softpooling layer with few parameters is applied to achieve the downsampling operation of the original branch. To be more specific, a 2×2 softpooling layer with the stride of 2 is incorporated into the transition branch of ECSPNet in the SDANet. In this way, it utilizes learnable parameters to assign weights to each output channel, preserving more comprehensive information from the original data. In addition, a channel shuffle layer is added after feature fusion to increase feature richness and diversity. Because the channel shuffle algorithm can split the channels of the feature map into several groups, shuffles the channels within each group, and then merges the different channels together. That considerably mitigate the problem of blocked information exchange between different groups caused by channel split, thereby improving the model's feature-sharing ability.

The overall process of information flow is described as follows. Firstly, the input feature tensor $I \in \mathbb{R}^{w \times h \times c}$ is separated into $I_1 \in \mathbb{R}^{w \times h \times c \times (1-s)}$ and $I_2 \in \mathbb{R}^{w \times h \times c \times s}$ along the channel direction according to the separation rate $s \in [0,1]$. Then, I_1 and I_2 are fused together to form a new feature tensor $I' \in \mathbb{R}^{w/2 \times h/2 \times c}$ after passing through two different branches (transition branch and dense block branch). Since the dense block branch contains more convolution operations than the transition branch, the parameters and FLOPs of the model lessen when s is set to a smaller value. Thirdly, the channel shuffle layer is added to disrupt the inter-channel information of I' and transmit output feature tensor to the next layer. The fundamental architecture of SDANet is composed of four stages with the proposed split-transform-merge strategy. Table 1 reveals the differences between Res2Net and the proposed SDANet from the aspects of structural components, parameters, and FLOPs.

4. Experimental results

All experimental results are performed on a Linux machine pre-installed with an Ubuntu 18.04 system. It is equipped with 4 T V100PCIe32GB GPUs, an Intel® Xeon® E5-2698 processor, and 256 GB of DDR4 RAM. To verify the classification effectiveness of the proposed network, several experiments are conducted on both ImageNet and CIFAR benchmarks in Section 4.1 and Section 4.2, respectively. Then, Section 4.3 shows our model's object detection capability based on the COCO dataset. After that, the SDANet effects of image segmentation are

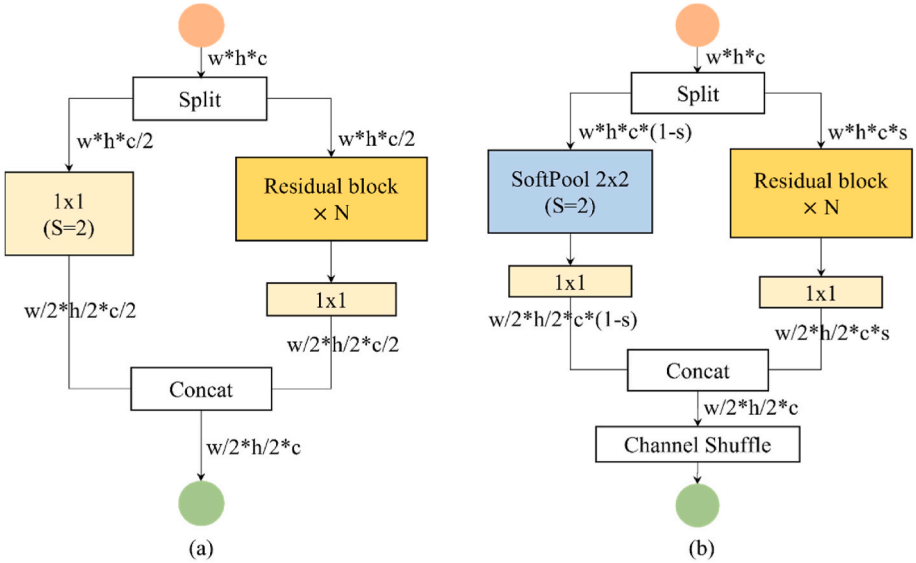

Fig. 4. Two different split-transform-merge strategies. (a) CSPNet; (b) ECSPNet.

Table 1
The comparison between the Res2Net-50 and SDANet-50 in terms of the model structure, parameters (#P) and FLOPs.

Stage	Output	Res2Net-50	SDANet-50	
	112×112	7×7, 64, stride 2		
		3×3, max pool, stride 2	2×2, softpool, stride 2	
1	56×56	$\begin{bmatrix} 1\times1, 128 \\ Res2[b=4], 128 \\ 1\times1, 256 \end{bmatrix} \times 3$	Split[s = 0.5]	
			$\begin{bmatrix} 3\times3[softpool], 64 \\ 1\times1, 128 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 64 \\ SDA[b=4], 64 \\ 1\times1, 128 \end{bmatrix} \times 3$
			Concat, 256	
2	28×28	$\begin{bmatrix} 1\times1, 256 \\ Res2[b=4], 256 \\ 1\times1, 512 \end{bmatrix} \times 4$	Split[s = 0.5]	
			$\begin{bmatrix} 3\times3[softpool], 128 \\ 1\times1, 256 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 128 \\ SDA[b=4], 128 \\ 1\times1, 256 \end{bmatrix} \times 4$
			Concat, 512	
3	14×14	$\begin{bmatrix} 1\times1, 512 \\ Res2[b=4], 512 \\ 1\times1, 1024 \end{bmatrix} \times 6$	Split[s = 0.5]	
			$\begin{bmatrix} 3\times3[softpool], 256 \\ 1\times1, 512 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 256 \\ SDA[b=4], 256 \\ 1\times1, 512 \end{bmatrix} \times 6$
			Concat, 1024	
4	7×7	$\begin{bmatrix} 1\times1, 1024 \\ Res2[b=4], 1024 \\ 1\times1, 2048 \end{bmatrix} \times 3$	Split[s = 0.5]	
			$\begin{bmatrix} 3\times3[softpool], 512 \\ 1\times1, 1024 \end{bmatrix}$	$\begin{bmatrix} 1\times1, 512 \\ SDA[b=4], 512 \\ 1\times1, 1024 \end{bmatrix} \times 3$
			Concat, 2048	
	1×1	global average pool, 1000-d fc, softmax		
#P		25.69M	18.05M	
FLOPs		4.28G	2.29G	

explored in Section 4.4. Finally, the ablation experiments are carried out in Section 4.5 in order to analyze and interpret the contributions of this study.

4.1. ImageNet classification

Firstly, the classification ability of the designed architecture is validated on the benchmark dataset ImageNet-1k (Russakovsky et al., 2015). ImageNet-1k is a large-scale dataset that includes 1000 classes, about 1.2 million images for training, 50,000 images for validation, and 100,000 images for testing. The challenges corresponding to this dataset consist of multiple tasks. In the image classification task, each model predicts a category label per image. The evaluation of the algorithm is completed by matching the predicted label with the ground truth (GT) label.

Based on the ImageNet classification dataset, performances of SOTA efficient models are summarized and visualized by comparing their respective accuracy, computational cost (FLOPs), and parameters (M). For this experiment, all the models in the comparison are set to the same training setup that is similar to (Gao et al., 2019a). For example, the experimental networks adopt SGD as the optimizer, and they are run for 100 training epochs. As illustrated in Fig. 5, the ShuffleNet is a light-weight model with little computation, but its classification accuracy is poor. On the contrary, the implemented SENet obtains the high accuracy of 80.02% at very large FLOPs (20.6G). The SDANet-101 achieves the best Top-1 accuracy of 81.24% with low FLOPs (3.03G) and few parameters (22.81M) by comparison with other networks.

Table 2 lists detailed results of some popular models on ImageNet-1k, which includes the referenced baselines cited from original papers and our implementations (the results of our implementations are filled in gray). According to the results in Table 2, there are three observations. Firstly, our proposed SDANets achieve significant gains of Top-1 error and Top-5 error with much complexity reduction. For example, SDANet-101 outperforms ResNet-101 by 3.65% of Top-1 error and 1.67% of Top-5 error, consuming less than about half of the parameters and GFLOPs. Compared to the latest multi-branch model (SKNet-101), the proposed SDANet-101 gains improvements for both classification errors, while it reduces 26.07M parameters and 5.47 GFLOPs. Moreover, SDANet-101 exceeds the recent advances of vision architectures in terms of a better trade-off between accuracy and complexity.

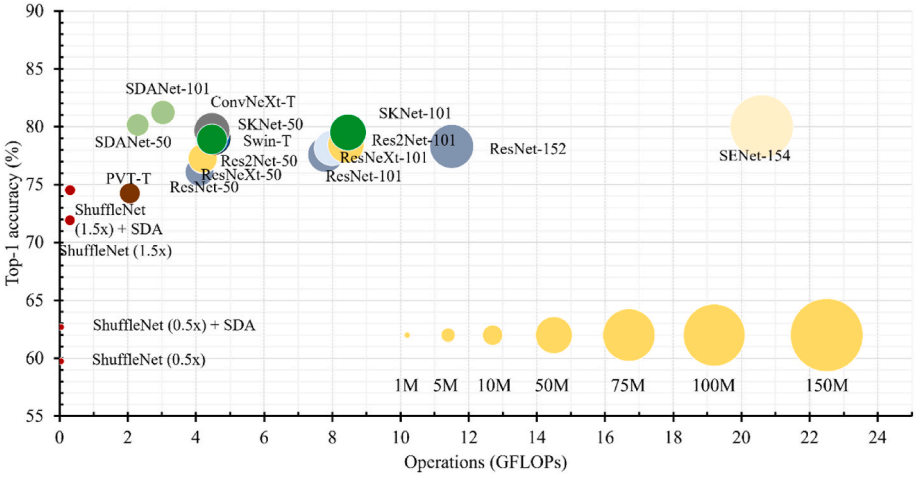

Fig. 5. Performance comparison of different models in terms of the ImageNet-1k accuracy, computation (GFLOPs), and model parameters (M).

Table 2
Comparison with recent SOTA architectures over classification accuracy and complexity. '#P' refers to the number of parameters. 'GFLOPs' represents the computations. 'Top-1/Top-5 err.' means the Top-1/Top-5 error rate. 'our impl.' means our implementation. 'images/sec' indicates the number of images per second.

Models	#P	GFLOPs	Top-1 err (%)	Top-5 err (%)	Images/sec
ResNet-50 (He et al., 2016)	–	3.8	22.85	6.71	–
ResNet-50 (our impl.)	25.5M	4.1	23.92	7.10	923.9
ResNeXt-50 (Xie et al., 2017)	–	–	22.20	–	–
ResNeXt-50 (our impl.)	25.03M	4.26	22.85	6.70	756.4
Res2Net-50 (Gao et al., 2019a)	–	4.2	22.01	6.15	–
Res2Net-50 (our impl.)	25.69M	4.28	22.73	6.64	794.3
SKNet-50 (Li et al., 2019)	27.50M	4.47	20.79	–	–
SKNet-50 (our impl.)	27.49M	4.52	21.08	5.81	514.5
SDANet-50 (ours)	**18.05M**	**2.29**	**19.84**	**5.06**	**961.2**
ResNet-101 (He et al., 2016)	–	7.60	21.75	6.05	–
ResNet-101 (our impl.)	44.50M	7.80	22.41	6.38	563.4
ResNeXt-101 (Xie et al., 2017)	–	–	21.20	5.60	–
ResNeXt-101 (our impl.)	44.18M	8.01	21.83	6.12	424.7
Res2Net-101 (Gao et al., 2019a)	–	–	20.81	5.57	–
Res2Net-101 (our impl.)	45.21M	8.10	21.54	6.10	451.7
SKNet-101 (Li et al., 2019)	48.90M	8.50	20.19	–	–
SKNet-101 (our impl.)	48.88M	8.55	20.48	5.53	285.1
PVT-S (Wang et al., 2021a)	24.50M	3.80	20.20	–	–
PVT-S (our impl.)	24.49M	3.82	21.75	6.31	684.1
ViT-S (Liu et al., 2022; Dosovitskiy, 2021)	22.00M	4.60	20.20	–	978.5
ViT-S (our impl.)	**22.74M**	4.68	21.43	6.34	**942.4**
Swin-T (Liu et al., 2021, 2022)	28.00M	4.50	18.70	–	757.9
Swin-T (our impl.)	28.50M	4.52	20.98	5.77	722.7
ConvNeXt-T (Liu et al., 2022)	29.00M	4.50	17.90	–	774.7
ConvNeXt-T (our impl.)	28.59M	4.47	20.33	5.53	738.9
SDANet-101 (ours)	22.81M	**3.03**	**18.76**	**4.71**	703.3

Moreover, some lightweight architectures are evaluated on ImageNet-1k in order to explore the performance gap among them. Since the complexity of the experimental networks in this section is low, their computations are expressed as MFLOPs instead of GFLOPs. As suggested in (Li et al., 2019), a typical model (ShuffleNetV2) with a lightweight and robust design is selected as a baseline to verify the SDA's generalization ability. Table 3 shows that SDA considerably reduces the baseline's classification error at different architectures scales. That demonstrates our proposed SDA module also performs well on the models with low complexity.

Table 3
Comparison with some lightweight architectures over classification accuracy and complexity. '#P' refers to the number of parameters. 'MFLOPs' represents the computations.

ShuffleNetV2	#P	MFLOPs	Top-1 err (%)
0.5 × (Ma et al., 2018)	1.4M	41	39.70
0.5 × (our impl.)	1.36M	42.5	40.26
0.5 × + SE	1.40M	42.7	38.56
0.5 × + SDA	1.37M	43.2	**37.31**
1.5 × (Ma et al., 2018)	3.5M	299	27.40
1.5 × (our impl.)	3.51M	304.5	28.08
1.5 × + SE	3.90M	305.6	26.53
1.5 × + SDA	3.58M	306.9	**25.47**

Since the proposed network is designed on the basis of the Res2Net model, the feature extraction capabilities of both models are measured via several images of Grad-CAM (Selvaraju et al., 2017). By computing the significance of spatial locations, CAM can highlight the influential area for each input image (Woo et al., 2018). Obviously, the CAM results of our SDANet50 have more accurate and concentrative maps for the objects with different scales. The strong feature representation ability at multi-scales makes the proposed network identify and localize objective regions precisely (see Fig. 6).

4.2. CIFAR classification

The two CIFAR datasets (Krizhevsky and Hinton, 2009) collected from independent search engines are used to assess the classification performance of SDANet on tiny images. The CIFAR-10 dataset is comprised of 60,000 small object-centric images in ten classes, with 32x32 pixels each. There are 50,000 training images and 10,000 testing images. Compared to CIFAR-10, the CIFAR-100 dataset divides the same images into 100 classes with more granular labelling. Each class has 500 images for training and 100 images for testing.

In this experiment, the architecture (Res2NeXt-29, 6c × 24w × 4b) in (Gao et al., 2019a) is considered as the primary reference due to its excellent performance. We first substitute the proposed SDANet block for the original Res2NeXt module and then add a new ECSPNet structure into each stage while remaining the other configurations unchanged. The Top-1 results of different models and their respective parameter numbers are shown in Table 4. In addition to the known results from recent papers, we also implement Res2NeXt-29 on CIFAR-10 for more comprehensive comparisons. The proposed SDANet exceeds the base model by 0.47% and 0.73% on CIFAR-10 and CIFAR-100, respectively. Among all the experimental models, our model achieves the lowest test error rate with the fewest parameters. Notably, SDANet obtains a significant accuracy gain than Wide ResNet with 62% fewer parameters, and it consistently suppresses the latest SOTA model (SKNet-29) on both CIFAR datasets with 50% fewer parameters.

Moreover, the classification performances of the model (Res2NeXt-29) and the proposed model (SDANet-29) on CIFAR datasets are illustrated in Fig. 7. By observing the validation results throughout the entire training process, the bold green curves fluctuate widely, whereas the bold red curves show a relatively smooth trend. That demonstrates our proposed SDANet has a more stable learning capability than Res2NeXt-29. In addition, the overall accuracy of SDANet is over the baseline for most of the epochs. Further, each plot has two conspicuous surge points at the 150th and 225th epochs, which is caused by the learning rate decay strategy. The specific change of the learning rate is determined by a pre-defined decay period and a multiplicative factor.

4.3. Object detection

The model's object detection capabilities were evaluated on the MS-COCO dataset (Lin et al., 2014), a vast image database comprising a multitude of objects and scenes. Specifically crafted for object detection tasks, it encompasses over 330,000 annotated images spanning more than 80 categories, offering an invaluable asset for research in the field of computer vision. The performances of the object detection task are reported on the MS-COCO dataset in Table 5. The purpose of this experiment is to demonstrate the superiority of the proposed model over the base model (Res2Net) from the perspective of the object detection. Both Cascade-RCNN (Cai and Vasconcelos, 2019) and VFNet (Zhang et al., 2021) are conducted as frameworks, and the backbone of Res2Net-50 versus SDANet-50 is replaced following the default settings of (Cai and Vasconcelos, 2019; Zhang et al., 2021). Regarding Average Precision (AP), the SDANet-50 surpasses the Res2Net-50 by 1.2% and 1.7% on Cascade RCNN and VFNet detectors, respectively. Also, the

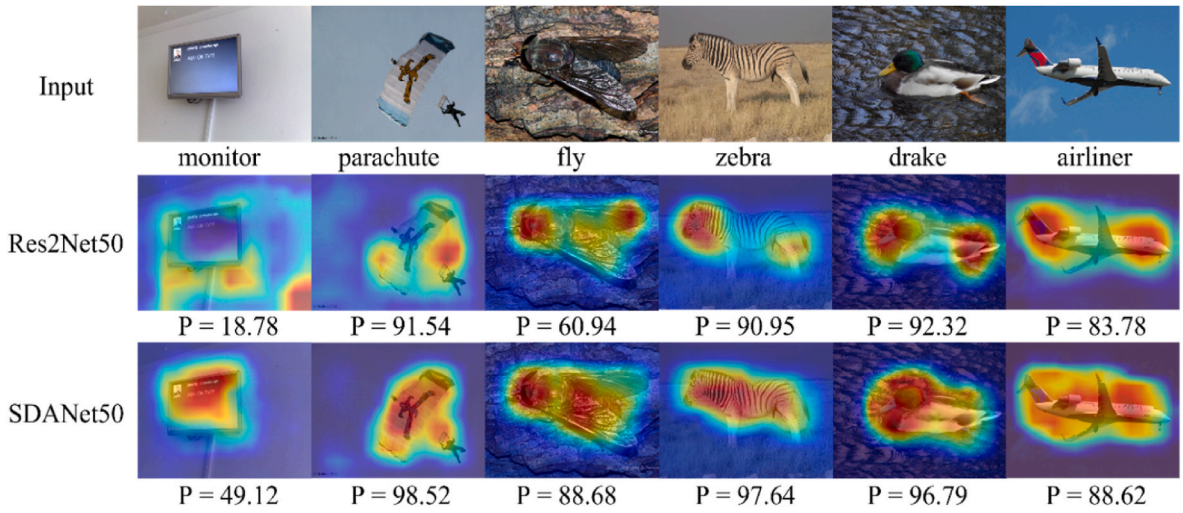

Fig. 6. Visualized examples of class activation mapping. The label of each input image is shown at the bottom of the first row, and 'P' means the specific softmax score for each category.

Table 4
Results of the models with different parameters on CIFAR datasets. 'w' is the basic width of conv3 × 3. 'c' is the number of groups. 'b' is the number of scales. 's' is the separation rate.

Model	#P	Top-1 err.	
		CIFAR-10	CIFAR-100
Wide ResNet (Zagoruyko and Komodakis, 2016)	36.5M	4.17	20.50
ResNeXt-29, 8c × 64w (Xie et al., 2017)	34.4M	3.65	17.77
ResNeXt-29, 16c × 64w (Xie et al., 2017)	68.1M	3.58	17.31
Res2NeXt-29, 6c × 24w × 4b (Gao et al., 2019a)	24.3M	3.73 (our impl.)	16.98
Res2NeXt-29, 6c × 24w × 4b-SE (Gao et al., 2019a)	26.0M	3.50 (our impl.)	16.68
SENet-29 (Hu et al., 2018)	35.0M	3.68	17.78
SKNet-29 (Li et al., 2019)	27.7M	3.47	17.33
SDANet-29, 24w × 4b, s = 0.5 (Proposed)	**13.9M**	**3.26**	**16.25**

SDANet-101 based model achieves 1.1% and 1.6% higher than Res2Net-101 based model. The experiment indicates the proposed backbone can boost the detection performances by embedding it into different detectors.

In addition, several examples of object detection results are presented in Fig. 8. The first row represents the results generated by the VFNet detector with our SDANet-101 backbone. The results obtained from the VFNet with Res2Net-101 backbone are in the second row. According to the images on the left, it implies that the Res2Net-101 has a lower confidence score for each object when the same objects are detected. The middle images illustrate that the SDANet-101 has better performance in detecting objects under an extremely dark environment. In the right images, Res2Net-101 fails to localize the person that is covered by others, yet it is successfully detected by SDANet-101. That demonstrates our proposed backbone is capable of catching any tiny and hidden objects thoroughly.

4.4. Image segmentation

In this section, image segmentation is explored from two distinct aspects, which include the semantic and instance segmentation tasks. With regard to the semantic segmentation, the performance of the proposed backbone is assessed and compared with other SOTA backbone models based on DeepLabv3+. This study uses two publicly available benchmark datasets to validate our model's multi-scale feature extraction ability. As suggested by the previous research (Gao et al., 2019a), the augmented PASCAL VOC12 dataset (Everingham et al., 2015) is considered one of the experimental datasets. A total of 12,031 images for 20 classes are separated into training and validation sets. Another one named Cityscapes (Cordts et al., 2016) is a semantic understanding image dataset of urban street scenes, which contains 2975 training images and 500 validation images. As shown in Table 6, the model with SDANet-50 surpasses the ResNet-50 based model by 2.6% on PASCAL VOC and 2.5% on Cityscapes. In addition, the proposed SDANet-101 achieves 81.5% Mean IoU on PASCAL VOC and 80.7% on Cityscapes, which is worthy of the best semantic segmentation results among all the experimental models.

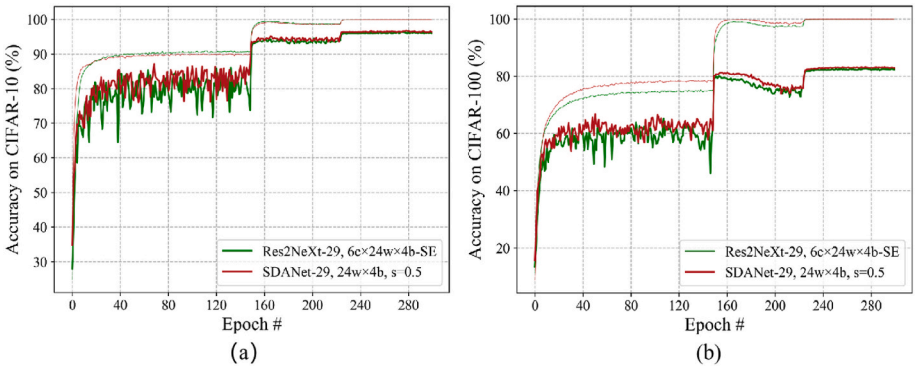

Fig. 7. The classification accuracy on CIFAR-10 (left) and CIFAR-100 (right). Thin and bold curves indicate training and validation accuracy respectively.

Table 5
Object detection results on the MS-COCO validation set. 'w' is the basic width of conv3 × 3. 'b' is the number of scales. 's' is the separation rate.

Method	Backbone	Setting	AP (%)	AP@IoU=0.5
Cascade RCNN (Cai and Vasconcelos, 2019)	Res2Net-50	26w × 4b	45.2	63.9
	SDANet-50	26w × 4b, s = 0.5	46.4	65.2
	Res2Net-101	26w × 4b	47.0	65.2
	SDANet-101	26w × 4b, s = 0.5	**48.1**	**67.0**
VFNet (Zhang et al., 2021)	Res2Net-50	26w × 4b	47.5	65.8
	SDANet-50	26w × 4b, s = 0.5	49.2	67.8
	Res2Net-101	26w × 4b	49.3	67.6
	SDANet-101	26w × 4b, s = 0.5	**50.9**	**69.4**

Furthermore, the proposed SDANet backbone is applied to the instance segmentation task based on the MS-COCO datasets. Table 7 tests the recent advanced methods, Cascade RCNN (Cai and Vasconcelos, 2019) and QueryInst (Fang et al., 2021), with Res2Net and SDANet as their backbones. Instance segmentation calculates the class probability, bounding box, and mask precision to acquire a more comprehensive and precise prediction. For Cascade RCNN, SDANet-50 exceeds its counterpart by 1.4% on box AP and 0.8% on mask AP, and SDANet-101 shows even more gains of 1.5% and 1.0%. For QueryInst, our backbone obtains better performance in both bounding box and mask predictions. That is because SDANet can extract more discriminant features in a broader range of RFs by using a new attention module and dilated convolutions.

4.5. Analysis and interpretation

In this section, an ablation study regarding the effectiveness of SDANet-29 with different settings on CIFAR-100 is conducted in Table 8. According to the first group of experiments, the efficacy of the proposed feature augmentation strategy is validated, with the incorporation of softpool and channel shuffle algorithms achieving a 1.11% decrease in terms of the model's Top-1 error rate. Additionally, the model with both multi-kernel structure and GANet attention mechanism reaches a remarkable result of Top-1 accuracy with relatively few parameters. Thus, all the experimental models in the second group are set to use the multi-kernel method and GANet attention. Results suggest that the

Table 6
Semantic segmentation results on the PASCAL VOC12 and Cityscapes datasets evaluated by the Mean intersection over union (IoU) metric.

Method	Backbone	Setting	Mean IoU (%)	
			PASCAL VOC	Cityscapes
DeepLabV3+ (Chen et al., 2018)	ResNet-50	64w	77.7	76.5
	Res2Net-50	26w × 4b	79.2	77.9
	SDANet-50	26w × 4b, s = 0.5	80.3	79.0
	ResNet-101	64w	79.0	78.1
	Res2Net-101	26w × 4b	80.2	79.6
	SDANet-101	26w × 4b, s = 0.5	**81.5**	**80.7**

Table 7
Instance segmentation results on MS-COCO dataset evaluated by the box average precision (box AP) and mask average precision (mask AP) metrics.

Method	Backbone	Setting	Box AP (%)	Mask AP (%)
Cascade RCNN (Cai and Vasconcelos, 2019)	Res2Net-50	26w × 4b	44.5	38.9
	SDANet-50	26w × 4b, S = 0.5	45.9	39.7
	Res2Net-101	26w × 4b	46.8	40.0
	SDANet-101	26w × 4b, S = 0.5	**48.3**	**41.0**
QueryInst (Fang et al., 2021)	Res2Net-50	26w × 4b	47.5	42.1
	SDANet-50	26w × 4b, S = 0.5	48.7	43.1
	Res2Net-101	26w × 4b	48.9	43.4
	SDANet-101	26w × 4b, S = 0.5	**50.2**	**44.3**

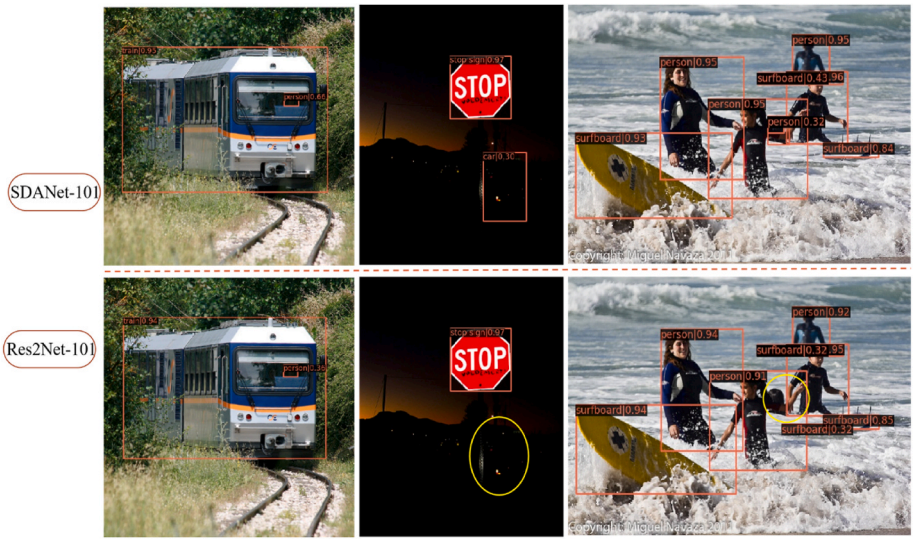

Fig. 8. Examples of object detection results (Images are from MS-COCO validation set) of two different backbone (Res2Net-101 and SDANet-101) using the VFNet detector.

Table 8
Results of SDANet-29 with different settings on the CIFAR-100 dataset.

Group	Model	Multi-kernel	GANet	Branch number (b)	Separation rate (s) $s \in [0, 1]$	Top-1 err. (%)	#P
1	SDANet-29 (b = 4, s = 0.5)	–	–	4	0.5	16.91	14.91M
		✓	–			16.48	**13.19M**
		–	✓			16.31	15.64M
		✓	✓			**16.25**	13.92M
	SDANet-29 (b = 4, s = 0.5, without feature augmentation strategy)	✓	✓			17.36	13.46M
2	SDANet-29 (Multi-kernel + GANet, s = 0.5)	✓	✓	2	0.5	17.86	**11.94M**
				4		16.25	13.92M
				6		**16.12**	16.03M
3	SDANet-29 (Multi-kernel + GANet, b = 4)	✓	✓	4	0.25	17.94	**13.75M**
					0.5	**16.25**	13.92M
					0.75	16.38	14.08M

optimum branch number for our proposed model is 4, which leads to a better trade-off between prediction accuracy and model complexity. In the last group, the separation rate becomes the only variable. The parameters of different models from s = 0.25 to s = 0.75 differ slightly, but the model with the separate rate of 0.5 achieves much higher accuracy than the model with s = 0.25 (the Top-1 error decreases from 17.94% to 16.25%). That demonstrates a suitable separate rate has a positive effect on improving the model's classification effectiveness.

5. Conclusion

This article introduces a lightweight yet efficient architecture, SDA, to enhance the multi-scale feature representation capability and improve all-around performances of classification, detection, semantic segmentation, and instance segmentation. The innovative residual-like multi-kernel method ensures refined feature extraction across various receptive fields, laying a robust foundation for subsequent processing. The introduction of a grouped attention mechanism further enriches the network by dynamically selecting and fusing features, optimizing the flow of information through the model. Additionally, the incorporation of a feature augmentation structure, powered by softpooling and channel shuffle functions, adds depth to the feature representation. The novel channel split and merge strategy, complemented by an adjustable coefficient, intelligently reduces computational overhead. These elements combine to form an SDA network that is not only highly effective but also seamlessly integrable with state-of-the-art models. Demonstrating a strong generalization ability, the SDANet has proven to deliver exceptional results across a multitude of datasets, solidifying its position as a valuable contribution to the field of computer vision. However, the present study is limit to the issue of interpretability, which indeed needs further investigation and resolution to improve the current model's comprehensibility and reliability. In the future, more effort should be put to explain the detailed learning processes, such as the feature extraction, weight optimization, and decision-making. We believe the presented architecture design can inspire and promote the future research for supervised deep learning.

CRediT authorship contribution statement

Hanxiang Wang: Writing – original draft, Methodology, Formal analysis. **Yanfen Li:** Writing – review & editing, Methodology, Data curation. **Tan N. Nguyen:** Formal analysis. **L. Minh Dang:** Conceptualization.

Declaration of competing interest

We have no conflicts of interest to disclose.

Data availability

Data will be made available on request.

References

Atif, S.M., Khan, S., Naseem, I., Togneri, R., Bennamoun, M., 2023. Multi-kernel fusion for RBF neural networks. Neural Process. Lett. 55 (2), 1045–1069.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Comput. Vis. Image Understand. 110 (3), 346–359.

Cai, Z., Vasconcelos, N., 2019. Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43 (5), 1483–1498.

Carreira, J., Madeira, H., Silva, J.G., 1998. Xception: a technique for the experimental evaluation of dependability in modern computers. IEEE Trans. Software Eng. 24 (2), 125–136.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV), pp. 801–818.

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z., 2020. Dynamic convolution: attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11030–11039.

Cordts, M., et al., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223.

Dosovitskiy, A., et al., 2021. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR).

de Lima, S.M., da Silva-Filho, A.G., Dos Santos, W.P., 2016. Detection and classification of masses in mammographic images in a multi-kernel approach. Comput. Methods Progr. Biomed. 134, 11–29.

Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. 111 (1), 98–136.

Fang, Y., et al., 2021. Instances as queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6910–6919.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P., 2019a. Res2net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 652–662.

Gao, Z., Wang, L., Wu, G., 2019b. Lip: local importance-based pooling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3355–3364.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M., 2019. Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558–567.

Howard, A.G., et al., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv Preprint arXiv:1704.04861.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Krizhevsky, A., Hinton, G., 2009. Learning Multiple Layers of Features from Tiny Images.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519.

Lin, T.-Y., et al., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, K., Moon, S., 2021. Dynamic parallel pyramid networks for scene recognition. IEEE Transact. Neural Networks Learn. Syst.

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

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.
Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 29.
Luo, Y., et al., 2022. CE-FPN: enhancing channel information for object detection. Multimed. Tool. Appl. 1–20.
Ma, N., Zhang, X., Zheng, H.-T., Sun, J., 2018. Shufflenet v2: practical guidelines for efficient cnn architecture design. In: Proceedings of the European Conference on Computer Vision. ECCV), pp. 116–131.
Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 1–66.
Park, J., Woo, S., Lee, J.-Y., Kweon, I.S., 2018. Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514.
Russakovsky, O., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv 1409–1556.
Stergiou, A., Poppe, R., Kalliatakis, G., 2021. Refining activation downsampling with SoftPool. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10357–10366.
Szegedy, C., et al., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H., 2020. CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391.
Wang, W., et al., 2021a. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578.
Wang, H., Li, Y., Dang, L.M., Lee, S., Moon, H., 2021b. Pixel-level tunnel crack segmentation using a weakly supervised annotation approach. Comput. Ind. 133, 103545.
Wang, H., Li, Y., Dang, L.M., Moon, H., 2022. An efficient attention module for instance segmentation network in pest monitoring. Comput. Electron. Agric. 195, 106853.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV), pp. 3–19.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500.
Xu, K., Li, D., Cassimatis, N., Wang, X., 2018. LCANet: end-to-end lipreading with cascaded attention-CTC. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, pp. 548–555.
Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv preprint arXiv: 1605.07146.
Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018a. Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision. ECCV), pp. 286–301.
Zhang, X., Zhou, X., Lin, M., Sun, J., 2018b. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856.
Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N., 2021. Varifocalnet: an iou-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8514–8523.
Zhang, H., et al., 2022. Resnest: split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2736–2746.
Zhou, D.-X., 2020. Theory of deep convolutional neural networks: downsampling. Neural Netw. 124, 319–327.