SCSA: Exploring the Synergistic Effects Between Spatial and Channel Attention

Yunzhong Si, Huiying Xu, Xinzhong Zhu, Wenhao Zhang, Yao Dong, Yuxing Chen, Hongbo Li

Abstract—Channel and spatial attentions have respectively brought significant improvements in extracting feature dependencies and spatial structure relations for various downstream vision tasks. While their combination is more beneficial for leveraging their individual strengths, the synergy between channel and spatial attentions has not been fully explored, lacking in fully harness the synergistic potential of multi-semantic information for feature guidance and mitigation of semantic disparities. Our study attempts to reveal the synergistic relationship between spatial and channel attention at multiple semantic levels, proposing a novel Spatial and Channel Synergistic Attention module (SCSA). Our SCSA consists of two parts: the Shareable Multi-Semantic Spatial Attention (SMSA) and the Progressive Channel-wise Self-Attention (PCSA). SMSA integrates multi-semantic information and utilizes a progressive compression strategy to inject discriminative spatial priors into PCSA's channel self-attention, effectively guiding channel recalibration. Additionally, the robust feature interactions based on the self-attention mechanism in PCSA further mitigate the disparities in multi-semantic information among different sub-features within SMSA. We conduct extensive experiments on seven benchmark datasets, including classification on ImageNet-1K, object detection on MSCOCO 2017, segmentation on ADE20K, and four other complex scene detection datasets. Our results demonstrate that our proposed SCSA not only surpasses the current state-of-the-art attention but also exhibits enhanced generalization capabilities across various task scenarios. The code and models are available at: https://github.com/HZAI-ZJNU/SCSA.

Index Terms—Multi-semantic information, semantic disparity, spatial attention, channel attention, synergistic effect.

I. INTRODUCTION

Attention mechanisms, by enhancing representations of interest, facilitate the learning of more discriminative features and are widely used in redistributing channel relationships and spatial dependencies. Existing universal attention methods can be primarily categorized into three types: channel attention [1]–[6], spatial attention [7]–[10], and hybrid channel-spatial

Yunzhong Si, Wenhao Zhang, Yao Dong and Yuxing Chen are with the College of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China. (e-mail: siyunzhong@zjnu.edu.cn; zwh2012918201@zjnu.edu.cn; dongyao@zjnu.edu.cn; cyx2001@zjnu.edu.cn).

Huiying Xu is with the College of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China, and also with the Research Institute of Hangzhou Artificial Intelligence, Zhejiang Normal University, Hangzhou 311231, China. (e-mail: xhy@zjnu.edu.cn).

Xinzhong Zhu is with the College of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China, the Research Institute of Hangzhou Artificial Intelligence, Zhejiang Normal University, Hangzhou 311231, China, and also with the Beijing Geekplus Technology Co., Ltd, Beijing, 100101, China. (e-mail: zxz@zjnu.edu.cn).

Hongbo Li is with the Beijing Geekplus Technology Co., Ltd, Beijing, 100101, China. (e-mail: jason.li@geekplus.com)

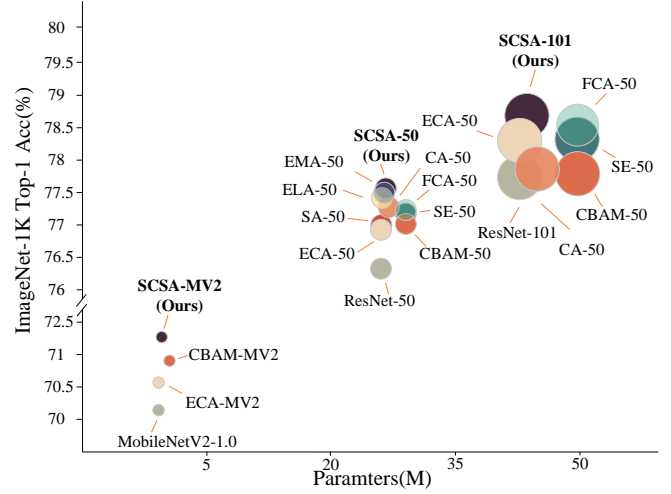Corresponding authors: Huiying Xu; Xinzhong Zhu.

Fig. 1. The accuracy of SCSA and various attention mechanisms is compared across multiple benchmark models on the ImageNet-1K validation set.

attention [11]–[18]. Their focuses differ: channel attention concentrates on extracting features of objects, while spatial attention is tailored to augment spatial information. We notice that spatial information represents semantic feature objects at the pixel level. Local spatial information captures low-semantic objects, while global spatial information perceives high-semantic objects. Thus, different spatial information, through their semantic associations, reflects distinct feature objects which are distributed across various features. To this end, we hypothesize that spatial information can guide channel learning. To further investigate guidance methodologies, we first review the current mainstream attention mechanisms.

CBAM [11] aggregates global spatial and channel information separately by chaining channel and spatial attention, but compressing all channel information leads to sharing across all spatial structures. This weakens the adaptability of spatial context to different feature maps. To overcome this, CPCA [19] introduces a channel-priority attention mechanism and depth-wise stripe convolutions, independently extracting spatial structures of each feature, significantly improving medical image segmentation. Furthermore, the EMA [17] module, based on grouped attention and cross-spatial multi-scale interactions, effectively integrates spatial information of both long and short-range dependencies but overlooks inter-group feature interactions. Although these hybrid attentions boost learning, they overlook the crucial guiding role of multi-semantic information in spatial-channel synergy.

How can we construct a synergistic mechanism where

spatial attention guides channel attention to enhance comprehensive learning, and channel attention modulates richer spatial-specific patterns from multi-semantic levels? Differing from the aforementioned methods, we explore the synergistic effects from three aspects: dimension decoupling, lightweight multi-semantic guidance, and semantic disparities mitigation, and propose a novel, plug-and-play Spatial and Channel Synergistic Attention(SCSA). Our SCSA is composed of a shareable Multi-Semantic Spatial Attention (SMSA) and a Progressive Channel Self-Attention (PCSA) linked sequentially. Our study initially employs multi-scale, depth-shared 1D convolutions to extract spatial information at various semantic levels from four independent sub-features. We utilize group normalization across four sub-features to hasten model convergence while avoiding the introduction of batch noise and the leakage of semantic information between sub-features. Subsequently, we input the SMSA-modulated feature maps into PCSA, incorporating progressive compression and channel-specific self-attention mechanisms (CSA). Our progressive compression strategy is designed to minimize computational complexity while preserving the spatial priors within SMSA, offering a practical trade-off. Moreover, our PCSA leverages an input-aware self-attention mechanism to effectively explore channel similarities, thereby mitigating semantic disparities among different sub-features in SMSA. We conducted extensive experiments across four visual tasks and seven benchmark datasets, illustrating the effectiveness of the multi-semantic synergy applied in our SCSA, and its superior generalization capability compared to other attention mechanisms. In summary, our contributions are as follows:

- We propose an efficient SMSA that utilizes multi-scale depth-shared 1D convolutions to capture multi-semantic spatial information for each feature channel, effectively integrating global contextual dependencies and multi-semantic spatial priors.
- Using SMSA, we develop feature-independent spatial structures and propose PCSA that calculates channel similarities and contributions guided by compressed spatial knowledge, mitigating semantic disparities in spatial structures.
- We connected SMSA and PCSA in series to create the SCSA, exploring synergistic effects through dimension decoupling, lightweight multi-semantic guidance, and semantic disparities mitigation. Our experiments confirm its superiority over current state-of-the-art attention mechanisms in various visual tasks and complex scenarios.

II. Related Work

A. Multi-Semantic Spatial Information

Multi-semantic spatial structures incorporate rich category and contextual information. Effectively integrating global context and local spatial priors enables models to learn higher-quality representations from various perspectives. The InceptionNets [20]–[23] pioneered a multi-branch approach, employing parallel vanilla convolutions of different sizes to capture varying receptive fields, significantly enhancing feature extraction capabilities. SKNet [2] incorporates multi-scale convolutions into channel attention, using the squeeze-and-excitation mechanism proposed by SENet [1] to integrate spatial priors with varying receptive fields. Benefiting from the global contextual modeling ability, ViT [10] employs MHSA to capture correlations at different spatial positions within distinct semantic sub-features, complemented by position embedding to compensate for spatial priors, achieving remarkable success in various downstream tasks. Currently, many studies develop efficient models [24]–[27] based on multi-semantic ideas, reducing parameters and computation for enhanced inference efficiency. Mamba [28] introduces a selectable state space model using scanning mechanisms and GPU parallelism to model global contextual dependencies with linear time complexity. Additionally, VMamba [29] proposes a cross-scanning module that extends 1D sequence scanning to 2D image scanning, effectively capturing multi-semantic global context information from four directions.

B. Attention Compression

Integrating attention mechanisms into various mainstream backbone or feature fusion networks enhances the model's understanding of fine-grained features and accuracy in feature representation. However, it inevitably leads to increased memory usage and computational time. CA [15] performs unidirectional spatial compression along the height (H) and width (W) dimensions separately, preserving spatial structures in one direction while aggregating global spatial information in another, mitigating information loss from global compression. SA [16] and EMA [5] reshape features into sub-features, reducing attention computation and parameters. However, reshape operations constrained by GPU bandwidth can lead to costly data transfers, with considerable time spent on data rearrangement, impacting training and inference speeds. CPCA [19] uses stripe convolutions in independent channels to reduce parameters in large-kernel convolutions. Recent studies also apply dimension decomposition in MHSA, with RMT [30] applying MHSA separately across H and W dimensions to minimize computational costs.

Although some of the aforementioned attention methods have proven effective in specific domains, they still suffer performance degradation when generalized to more complex scenarios.

III. Method

In this section, we begin by discussing the SMSA module, which explores the benefits of lightweight multi-semantic information guidance. Next, we introduce the PCSA module, which utilizes a progressive compression strategy and channel-wise self-attention to mitigate semantic disparities. The synergistic effects of multi-semantic guidance and semantic disparities mitigation motivate us to propose SCSA module. The overall architecture is shown in Figure 2.

A. Shared Multi-Semantic Spatial Attention

1) Spatial and Channel Decomposition: Decomposition techniques in neural network architectures significantly reduce the number of parameters and computational cost. Inspired by
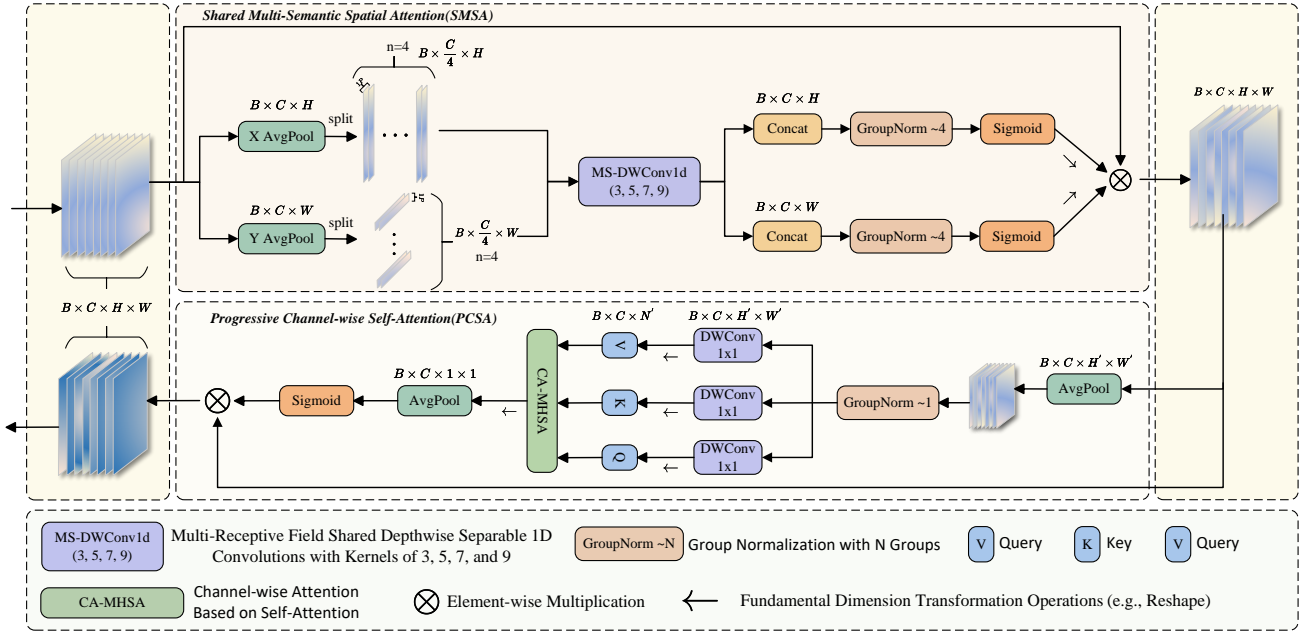
Fig. 2. An illustration of our proposed SCSA, which uses multi-semantic spatial information to guide the learning of channel-wise self-attention. B denotes the batch size, C signifies the number of channels, and H and W correspond to the height and width of the feature maps, respectively. The variable n represents the number of groups into which sub-features are divided, and $1P$ denotes a single pixel.

TABLE I
COMPARISON OF OUR METHOD, BASED ON THE UPERNET MODEL, WITH OTHER ATTENTION MECHANISMS FOR SEMANTIC SEGMENTATION PERFORMANCE ON THE ADE20K BENCHMARK.

Methods	UperNet		
	Params(M)	FLOPs(G)	mIoU(%)
ResNet-50	64.10	1895	40.20
+ CBAM	66.62	1895	39.62
+ CPCA	65.94	1927	39.68
+ SE	66.62	1895	39.94
+ SA	64.10	1895	40.01
+ ECA	64.10	1895	40.46
+ FCA	66.61	1895	41.09
+ SCSA(Ours)	64.16	1895	**41.14**
ResNet-101	83.09	2051	42.74
+ CBAM	87.84	2051	41.65
+ ECA	83.09	2051	42.63
+ SE	87.84	2051	42.66
+ FCA	87.83	2051	43.22
+ SCSA(Ours)	83.22	2051	**43.76**

TABLE II
COMPARISON OF OUR METHOD, BASED ON THE MASK R-CNN, WITH OTHER ATTENTION MECHANISMS FOR INSTANCE SEGMENTATION PERFORMANCE ON MSCOCO VAL2017.

Methods	Mask R-CNN					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50	34.8	55.9	36.9	16.4	37.4	50.2
+ CBAM	35.4	56.9	37.6	17.4	38.3	50.6
+ ECA	35.5	57.6	37.6	16.6	38.4	**52.0**
+ FCA	35.5	57.2	37.6	17.1	38.6	51.3
+ SE	35.7	57.3	38.1	**17.7**	38.6	50.9
+ SA	35.7	57.7	38.0	17.2	38.7	51.5
+ CA	35.8	57.5	38.2	16.9	38.5	51.7
+ SCSA	**36.1**	**58.4**	**38.3**	17.2	**39.1**	51.9

is presented as follows:

$$X_H^i = X_H[:, (i-1) \times \frac{C}{K} : i \times \frac{C}{K}, :] \quad (1)$$

$$X_W^i = X_W[:, (i-1) \times \frac{C}{K} : i \times \frac{C}{K}, :] \quad (2)$$

X^i represents the i-th sub-feature, where $i \in [1, K]$. Each sub-feature is independent, facilitating efficient extraction of multi-semantic spatial information.

2) Lightweight Convolution Strategies Across Disjoint Sub-features: After partitioning the feature set into exclusive sub-features, we aim to efficiently capture distinct semantic spatial structures within each sub-feature. Inspired by extensive research [25], [26], [31] on reducing feature redundancy, which reveal that such redundancy is likely due to intense interactions among features, we also observe varied spatial structures among features. Based on these insight, and aiming to enrich semantic information, enhance semantic coherence, and minimize semantic gaps, we apply depth-wise 1D convolutions with kernel sizes of 3, 5, 7, and 9 in four sub-features. Furthermore, to address the limited receptive field

the structure of 1D sequences in NLP tasks, in our study, we decompose the given input $X \in \mathbb{R}^{B \times C \times H \times W}$ along the height and width dimensions. We apply global average pooling to each dimension, thereby creating two unidirectional 1D sequence structures: $X_H \in \mathbb{R}^{B \times C \times W}$ and $X_W \in \mathbb{R}^{B \times C \times H}$. To learn varying spatial distributions and contextual relationships, we partition the feature set into K identically sized, independent sub-features, X_H^i and X_W^i, with each sub-feature having a channel count of $\frac{C}{K}$. In this paper, we set the default value $K = 4$. The process of decomposing into sub-features

caused by decomposing features into H and W dimensions and applying 1D convolutions separately, we use lightweight shared convolutions for alignment, implicitly modeling the dependency between the two dimensions by learning consistent features in both dimensions. The ablation details regarding them are provided in Table III. The implementation process for extracting multi-semantic spatial information is defined as follows:

$$\tilde{X}_H^i = DWConv1d_{k_i}^{\frac{C}{K} \to \frac{C}{K}}(X_H^i) \tag{3}$$

$$\tilde{X}_W^i = DWConv1d_{k_i}^{\frac{C}{K} \to \frac{C}{K}}(X_W^i) \tag{4}$$

$\tilde{X}^i$ represents the spatial structural information of the i-th sub-feature obtained after lightweight convolutional operations. k_i denotes the convolution kernel applied to the i-th sub-feature.

To accurately compute spatial attention maps for each feature, we aggregate distinct semantic sub-features and use Group Normalization (GN) with K groups for normalization. We opt for GN over the conventional Batch Normalization (BN) because our study finds that GN is superior in distinguishing semantic differences among sub-features. GN allows for the independent normalization of each sub-feature without introducing batch statistical noise, effectively mitigating semantic interference between sub-features and preventing attention dilution. This approach is validated by ablation studies shown in Table III. Finally, spatial attention is generated using a simple Sigmoid activation function, which activates and suppresses specific spatial regions. The computation of output features is as follows:

$$Attn_H = \sigma(GN_H^K(Concat(\tilde{X}_H^1, \tilde{X}_H^2, ..., \tilde{X}_H^K))) \tag{5}$$

$$Attn_W = \sigma(GN_W^K(Concat(\tilde{X}_W^1, \tilde{X}_W^2, ..., \tilde{X}_W^K))) \tag{6}$$

$$X_s = Attn_H \times Attn_W \times X \tag{7}$$

$\sigma(\cdot)$ denotes the Sigmoid normalization, while $GN_H^K(\cdot)$ and $GN_W^K(\cdot)$ represent GN with K groups along the H and W dimensions, respectively.

B. Progressive Channel-wise Self-Attention

A prevalent approach to compute channel attention is through convolutional operations that explore dependencies among channels. There is limited theoretical research on how convolutional layers explicitly model these dependencies, aside from achieving it through backpropagation and gradient updates. Inspired by the significant advantages of the ViT [10] in utilizing MHSA for modeling similarities among different tokens in spatial attention calculations, we propose combining the MHSA concept with modulated spatial priors from SMSA to compute inter-channel similarities. Moreover, to preserve and utilize the multi-semantic spatial information extracted by SMSA, and to reduce the computational cost of MHSA, we employ a progressive compression method, which reflects the **guidance** in our synergistic effects. Compared to traditional convolutional operations, PCSA exhibits stronger input perception capabilities and effectively utilizes the spatial priors provided by SMSA to deepen learning. The implementation details of our PCSA are as follows:

$$X_p = Pool_{(7,7)}^{(H,W) \to (H',W')}(X_s) \tag{8}$$

$$F_{proj} = DWConv1d_{(1,1)}^{C \to C} \tag{9}$$

$$Q = F_{proj}^Q(X_p), K = F_{proj}^K(X_p), V = F_{proj}^V(X_p) \tag{10}$$

$$X_{attn} = Attn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{C}})V \tag{11}$$

$$X_c = X_s \times \sigma(Pool_{(H',W')}^{(H',W') \to (1,1)}(X_{attn})) \tag{12}$$

TABLE III
ABLATION STUDIES ON THE DESIGN STRATEGY OF SCSA, CONDUCTED AT A 224x224 RESOLUTION, USING THE IMAGENET-1K VALIDATION SET. THE ABBREVIATION "PC" DENOTES PROGRESSIVE COMPRESSION. $Gi(K_1, K_2, \ldots, K_i)$ DENOTES SPLITTING X INTO i SUB-FEATURES AND APPLYING A 1D CONVOLUTION OF SIZE K_i TO EACH i-TH SUB-FEATURE.

Ablations	Variants	Throughput (imgs/s)	Top-1 (%)
Baseline	**SCSA-50**	2009	**77.49**
Macro Design	w/o SMSA	2217	76.72
	w/o PCSA	2155	77.44
	w/o PC	1982	77.31
	w/ Multi-head + Shuffle	2082	77.35
Ordering	PCSA Prior	2005	77.20
	GN Prior	2010	77.47
Micro Design	GN→BN	1999	77.19
	Shared → Unshared	1981	77.32
	Scaled: $\sqrt{C} \to \sqrt{H*W}$	2001	77.34
Branch	G1(3)	2085	77.24
	G1(7)	2063	77.17
	G2(3,7)	2040	77.32

$Pool_{(7,7)}^{(H,W) \to (H',W')}(\cdot)$ denotes a pooling operation with a kernel size of 7x7 that rescales the resolution from (H, W) to (H', W'). $F_{proj}(\cdot)$ represents the mapping function that generates the query, key, and value.

It's important to note that, unlike the MHSA in the ViTs where $Q, K, V \in \mathbb{R}^{B \times N \times C}$ with $N = HW$, in our PCSA's CA-MHSA, self-attention is computed along the channel dimension, with $Q, K, V \in \mathbb{R}^{B \times C \times N}$.

C. Synergistic Effects

The synergistic spatial and channel attention mechanisms aim to complement each other. In our work, we propose a novel concept of guiding channel attention learning through spatial attention. Drawing from the connection methods of CBAM [11] and CPCA [19], we employ a simple serial connection to integrate our SMSA and PSCA modules. Our innovation lies in the meticulously designed spatial and channel attentions: spatial attention extracts multi-semantic spatial information from each feature, providing precise spatial priors for channel attention computation; channel attention refines the semantic understanding of local sub-feature X^i by leveraging the overall feature map X, mitigating semantic disparities caused by multi-scale convolution in SMSA. Additionally, unlike previous approaches [1], [11], [12], [15], we do not employ channel compression, effectively preventing the loss of crucial features. Ultimately, our constructed SCSA is as follows:

$$SCSA(X) = PCSA(SMSA(X)) \tag{13}$$

IV. EXPERIMENTS

In this section, we first introduce the experimental details. Next, we conduct experiments on four visual tasks, comparing our proposed SCSA with other state-of-the-art attention mechanisms. Following this, in Section IV-E, we perform a comprehensive ablation study on our meticulously designed SCSA from four different perspectives.

TABLE IV
COMPARISON OF OUR PROPOSED SCSA WITH OTHER STATE-OF-THE-ART ATTENTION MECHANISMS ACROSS MULTIPLE BENCHMARK MODELS AT A 224x224 RESOLUTION ON THE IMAGENET-1K VALIDATION SET.

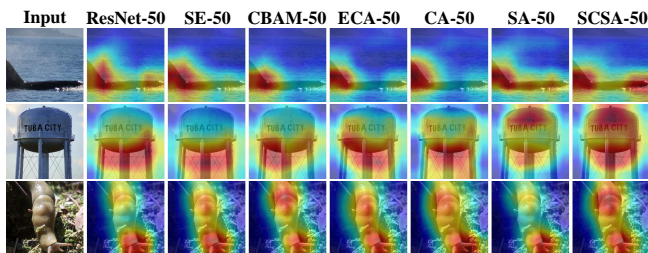| Backbones	Type	Methods	Params(M)	FLOPs(G)	Throughput(imgs/s)	Top-1(%)	Top-5(%)
ResNet-50	–	ResNet [32]	25.56	4.02	2433	76.39	93.09
	Channel	ECANet [5]	25.56	4.11	2109	77.05	93.43
		SENet [1]	28.07	4.11	2077	77.23	93.56
		FcaNet [3]	28.07	4.11	1905	77.29	93.64
	Hybrid	CPCA [19]	27.40	4.87	1379	75.80	92.57
		CBAM [11]	28.07	4.12	1687	77.12	93.50
		SANet [16]	25.56	4.12	1493	77.12	93.64
		ELA [18]	25.59	4.11	2233	77.25	93.52
		CA [15]	25.69	4.11	2244	77.37	93.52
		EMA [17]	25.57	4.18	1861	77.43	**93.79**
		SCSA(Ours)	25.62	4.12	2019	**77.49**	93.60
ResNet-101	–	ResNet [32]	44.55	7.83	1588	77.76	93.81
	Channel	ECANet [5]	44.55	7.83	1408	78.32	93.99
		SENet [1]	49.30	7.84	1399	78.40	94.05
		FcaNet [3]	49.29	7.84	1242	78.51	94.10
	Hybrid	CBAM [11]	49.30	7.84	1118	78.09	94.07
		CA [15]	44.80	7.84	1437	78.11	93.92
		SCSA(Ours)	44.68	7.85	1298	**78.56**	**94.31**
MobileNetV2-1.0	–	MobileNetV2 [33]	3.51	0.31	6693	71.54	90.11
	Channel	ECANet [5]	3.51	0.31	5746	72.02	90.35
	Hybrid	CBAM [11]	4.07	0.32	4539	72.43	90.49
		SCSA(Ours)	3.63	0.34	2751	**72.72**	**90.81**

Fig. 3. Comparative attention visualizations for 'layer 4.2' across multiple models, generated using samples randomly selected from different categories of the ImageNet-1K validation set, through Grad-CAM [34].

A. Implementation Details

To evaluate our proposed SCSA on ImageNet-1K [35], we select three mainstream backbone networks based on CNN architectures, including ResNet-50, ResNet-101 [32] and MobileNetV2-1.0 [33]. Specifically, for models based on ResNets, we employ an SGD optimizer with a momentum of 0.9, a weight decay of 1e-4, and an initial learning rate of 0.05. which is reduced tenfold every 30 epochs. The models are trained using a batch size of 128 over 100 training epochs. For training MobileNetV2 with our SCSA, we follow the settings used in ECANet [5], the optimizer has a momentum of 0.9 and a weight decay of 4e-5, with an initial learning rate of 0.045. This rate decrease linearly by a factor of 0.98, and the batch size is set to 96 for 400 training epochs. Notably, to enhance training efficiency, we employ Automatically Mixed Precision(AMP) training on a single NVIDIA RTX 4090 GPU for the classification tasks.

We evaluate our SCSA on MSCOCO 2017 [36] using Faster R-CNN [37], Mask R-CNN [38], Cascade R-CNN [39], and RetinaNet [40]. These detectors are implemented using the MMDetection [41] toolboxes with default settings. Input images are scaled proportionally by their shorter side to 800. All models are trained using an SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4, with a batch size of 2 per GPU, over a total of 12 epochs. Faster R-CNN, Mask R-CNN and Cascade R-CNN started with a learning rate of 0.0025, while RetinaNet starts at 0.00125. The learning rates for all models are decreased by a factor of 10 at the 8th and 11th epochs. We fine-tuned the model on the MSCOCO train2017 dataset for 12 epochs using a single NVIDIA H800 GPU and reported comparative results on val2017.

We further validate our method on ADE20K [42] with the UperNet [43] for semantic segmentation. Following common practices [44], [45], we also utilize the MMSegmentation [46] toolboxes, set the batch size to 16, and conduct 80k training iterations. All models are trained using an SGD optimizer with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 5e-4. We also conduct training and inference using a single NVIDIA H800 GPU.

All models are trained with the default random seed 0.

B. Image Classification

We compare our SCSA against other state-of-the-art attention mechanisms such as SENet [1], CBAM [11], ECANet [5], FcaNet [3], CA [15], SANet [16], EMA [17], CPCA [19], and ELA [18]. As shown in Figure 1 and Table IV, our SCSA achieved the highest Top-1 accuracy across networks of different scales, with negligible parameter count and computational complexity. Within hybrid architectures, our

TABLE V
Comparison of the performance of different attention mechanisms for object detection on MSCOCO val2017, utilizing models such as Faster R-CNN, Cascade R-CNN, and RetinaNet. All models were fine-tuned using the "1×" schedule.

Detectors	Methods	Params(M)	FLOPs(G)	AP(%)	AP_{50}(%)	AP_{75}(%)	AP_S(%)	AP_M(%)	AP_L(%)
Faster R-CNN	ResNet-50	41.75	187.20	37.6	58.7	40.9	21.5	41.2	48.1
	+ FCA	44.27	187.31	38.4	59.8	41.5	22.8	42.4	48.9
	+ ECA	41.75	187.20	38.5	60.0	41.4	22.6	42.6	49.4
	+ SE	44.27	187.20	38.7	60.2	41.6	23.2	42.4	49.3
	+ CA	41.89	187.21	39.0	60.6	42.3	23.2	42.8	49.5
	+ SCSA(Ours)	41.81	187.35	**39.3**	**60.6**	**42.8**	**23.2**	**43.1**	**50.2**
	ResNet-101	60.75	255.43	40.2	61.3	43.8	23.9	44.2	51.8
	+ FCA	65.49	255.60	40.6	62.2	44.1	23.8	44.9	52.5
	+ SE	65.49	255.44	40.8	62.2	44.4	**24.9**	44.7	53.0
	+ ECA	60.75	255.43	40.9	62.4	44.3	24.2	45.0	53.0
	+ CA	60.99	255.45	41.1	62.2	44.8	24.1	45.0	53.5
	+ SCSA(Ours)	60.88	255.74	**41.5**	**62.9**	**45.4**	24.6	**45.3**	**53.7**
Cascade R-CNN	ResNet-50	69.40	214.84	40.3	58.9	43.8	22.5	43.8	52.8
	+ FCA	71.91	214.94	41.3	60.2	44.6	24.1	44.9	53.7
	+ SE	71.91	214.84	41.4	60.2	44.9	24.5	44.7	54.0
	+ CBAM	71.91	214.87	41.4	60.2	45.0	24.5	44.6	54.3
	+ ECA	69.40	214.84	41.7	60.7	45.2	**24.7**	45.4	54.3
	+ SCSA(Ours)	69.46	214.99	**42.1**	**61.4**	**45.7**	24.6	**45.5**	**54.3**
	ResNet-101	88.39	283.07	42.6	61.1	46.6	24.9	46.7	55.7
	+ SE	93.13	283.08	43.2	62.3	47.2	25.8	47.1	56.2
	+ FCA	93.13	283.24	43.4	62.5	47.6	25.5	47.3	56.8
	+ ECA	88.39	283.07	43.7	62.7	47.5	25.5	47.7	56.8
	+ CA	88.64	283.09	43.8	62.8	48.0	26.0	47.6	57.4
	+ SCSA(Ours)	88.52	283.38	**44.2**	**63.1**	**48.2**	**26.0**	**48.2**	**57.5**
RetinaNet	ResNet-50	37.97	214.68	36.5	55.5	39.1	20.2	40.1	48.1
	+ FCA	40.48	214.78	37.3	57.2	39.3	21.6	40.9	49.0
	+ SE	40.49	214.68	37.4	57.0	40.0	21.5	41.3	49.0
	+ ECA	37.97	214.68	37.5	57.2	39.8	21.5	41.1	49.5
	+ CBAM	40.49	214.71	37.6	57.0	40.2	22.0	**41.6**	48.7
	+ SCSA(Ours)	38.03	214.83	**37.9**	**57.6**	**40.2**	**22.5**	41.3	**49.7**
	ResNet-101	56.96	282.91	39.3	58.7	41.9	22.8	43.5	51.8
	+ SE	61.71	282.92	39.8	59.9	42.2	22.9	43.8	52.1
	+ FCA	61.70	283.08	39.9	60.0	42.4	22.9	**44.6**	52.4
	+ CA	57.21	282.93	40.2	60.0	43.0	23.2	44.3	52.8
	+ ECA	56.96	282.91	40.3	60.4	42.9	23.4	44.2	52.7
	+ SCSA(Ours)	57.09	283.22	**40.5**	**60.8**	**43.6**	**23.7**	44.3	**53.1**

method's inference speed based on ResNet is second only to CA, but it offer a better balance of accuracy, speed, and model complexity with a moderate model width.

C. Object Detection

We evaluate the generalization ability of our SCSA on general detection tasks to verify its effectiveness in enhancing feature extraction. We use ResNet-50 and ResNet-101 as the backbone networks and FPN [47] as the feature fusion sub-network.

As shown in Table V, our method outperforms other state-of-the-art attention methods across various detectors, model sizes, and object scales. For Faster R-CNN, our SCSA improves by 1.7% and 1.3% compared to the original ResNet-50 and ResNet-101, respectively. We also provide visualizations based on MSCOCO val2017 in Appendix C and detailed experiments on more complex scenarios (e.g., small targets on the VisDrone2019 [48] dataset, dark environments on the ExDark [49] dataset, and infrared scenes on the FLIR-ADAS-V2 [50] dataset) in Appendix E. These further demonstrating the effectiveness and generalizability of our method.

D. Segmentation

We also test its performance in semantic segmentation on ADE20K and instance segmentation on MSCOCO 2017. Some methods (e.g., CBAM, CPCA, SE, SA, ECA) have not previously been tested on ADE20K for semantic segmentation, so we conduct extensive comparative experiments based on the UperNet [43] network. As shown in Tables I and II, our SCSA significantly outperforms other attention methods. Specifically, SCSA improves performance by 0.94% and 1.02% on ResNet-50 and ResNet-101, respectively, while other methods only achieve improvements of 0.1% to 0.2%, and some even fall below the baseline model. These results demonstrate that our method, based on multi-semantic spatial information, performs exceptionally well in pixel-level tasks.

E. Ablation Study

As shown in Table III, we use SCSA-50 as a baseline on ImageNet-1K for ablations across four aspects.

1) Macro Design: We separately validate the SMSA and PCSA modules. Both demonstrate significant accuracy improvements, with SMSA notably enhancing classification accuracy by 1.05%. Without progressive compression in PCSA,

accuracy drops by 0.18%, and this is primarily because the channel attention mechanism cannot utilize the discriminative spatial priors provided by SMSA for its computations. Additionally, when replacing the single attention head in PCSA with multi-head and channel shuffle operation, performance decrease from 77.49% to 77.35%. This phenomenon is primarily attributed to the strong inter-channel interactions facilitated by the single head, which effectively alleviate semantic disparities observed in SMSA.

2) Ordering: When PCSA is moved ahead of SMSA, there is a 0.29% drop in Top-1 accuracy. This decrease is attributed to the early channel attention, where spatial information across features remains unmodulated by spatial attention, lacking precise spatial priors to guide channel recalibration. This effectively validates our hypothesis that spatial attention can guide channel learning. Concerning the timing of normalization, placing GN before the attention computation leads to a slight reduction in accuracy. This could be due to pre-normalization diminishing the distinct semantic patterns inherent to different sub-features, thus impacting the representation of diverse semantic features.

3) Micro Design: Replacing GN with BN results in a decrease in both accuracy and inference speed, with the Top-1 accuracy dropping from 77.49% to 77.19%. This decline is attributed to GN's superior ability to preserve the independence of semantic patterns among sub-features, thereby minimizing semantic interference. Conversely, BN is highly sensitive to batch size, introducing additional batch statistical noise for multi-semantic information. These insights suggest that GN may be a more suitable choice in convolution layers that involve multiple semantics. Furthermore, the decline in accuracy and increase in parameters with unshared convolutions further validate the effectiveness of using shared convolutions to consistently learn and model features and dependencies across the H and W dimensions.

4) Branch: Reducing the diversity of convolution kernels led to a decrease in accuracy, underscoring the significant impact of multi-semantic information on enhancing feature extraction capabilities. Drawing on the synergistic approach proposed in our SCSA, we leverage semantic space information to bolster feature extraction and employ channel self-attention to facilitate semantic interaction, effectively mitigating disparities across different semantic spaces and promoting better fusion of multi-semantic information.

V. ANALYSIS

A. Visualization of Attention

As shown in Figure 3, our SCSA distinctly focuses on multiple key regions under similar receptive field conditions, significantly minimizing critical information loss, and providing rich feature information for the ultimate downstream tasks. This advantage arises from the SCSA's synergistic design, which preserves critical information in both spatial and channel domains during their attention computations, further emphasizing its superior representational capabilities compared to other attention mechanisms. Detailed analysis and visualization results for detection and segmentation tasks can be found in Appendix.

VI. CONCLUSION

In this paper, we explore the synergy between spatial and channel dimensions through dimension decoupling, lightweight multi-semantic guidance, and semantic disparities mitigation, proposing SCSA—a novel, plug-and-play spatial and channel synergistic attention mechanism. Extensive experiments across multiple visual tasks demonstrate the enhanced performance and robust generalization capabilities of our SCSA compared to other state-of-the-art attention mechanism. We hope that our research will facilitate the exploration of synergistic properties across multiple dimensions in various domains.

VII. ACKNOWLEDGMENTS

REFERENCES

[1] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[2] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.
[3] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.
[4] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 794–11 803.
[5] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.
[6] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 096–10 105.
[7] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018.
[8] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[9] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
[12] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.
[13] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.
[14] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.
[15] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[16] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2235–2239.

[17] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] W. Xu and Y. Wan, "Ela: Efficient local attention for deep convolutional neural networks," *arXiv preprint arXiv:2403.01123*, 2024.

[19] H. Huang, Z. Chen, Y. Zou, M. Lu, and C. Chen, "Channel prior convolutional attention for medical image segmentation," *arXiv preprint arXiv:2306.05196*, 2023.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[23] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[24] X. Zhang, X. Zhou, M. Lin, and J. 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.

[25] N. Ma, X. Zhang, H.-T. Zheng, and J. 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] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12021–12031.

[27] A. Wang, H. Chen, Z. Lin, H. Pu, and G. Ding, "Repvit: Revisiting mobile cnn from vit perspective," *arXiv preprint arXiv:2307.09283*, 2023.

[28] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[29] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[30] Q. Fan, H. Huang, M. Chen, H. Liu, and R. He, "Rmt: Retentive networks meet vision transformers," *arXiv preprint arXiv:2309.11523*, 2023.

[31] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580–1589.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. 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.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[39] Z. Cai and N. 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.

[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[41] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[42] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[43] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[44] W. Yu, P. Zhou, S. Yan, and X. Wang, "Inceptionnext: When inception meets convnext," *arXiv preprint arXiv:2303.16900*, 2023.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[46] M. Contributors, "Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[48] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[49] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[50] T. FLIR, "Free teledyne flir thermal dataset for algorithm training," 2023, https://www.flir.com/oem/adas/adas-dataset-form/.

[51] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

APPENDIX A
INTEGRATING SCSA INTO THE BLOCK

Figure 4 shows our SCSA module integrated into blocks of ResNet and MobileNetV2. In our work, for ResNet, we consider the appropriate model width, which differs slightly from other attention mechanisms. For MobileNetV2, we account for the DWConv's need to compensate for its lack of expressive capability by increasing the model width. Therefore, we place it in the same position as other attention mechanisms.

APPENDIX B
VISUALIZATION OF SEGMENTATION RESULTS

A. Instance Segmentation

To validate the effectiveness of incorporating synergistic multi-semantic information into SCSA, we select three random samples from MSCOCO val2017 for instance segmentation with Mask R-CNN. As shown in Figure 5, our method segments obscured and overlapping objects more comprehensively and accurately, achieving higher confidence scores. These results underscore the benefits of our method in
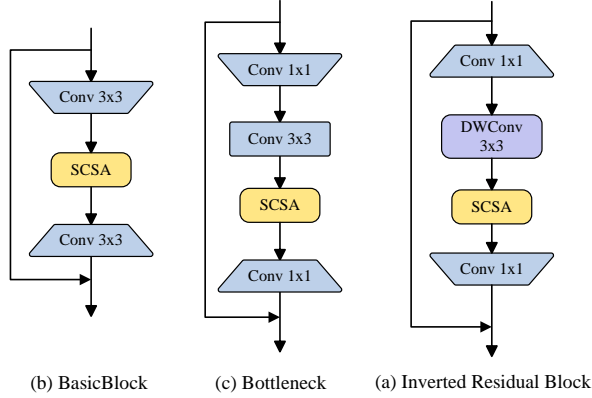
(b) BasicBlock   (c) Bottleneck   (a) Inverted Residual Block

Fig. 4. Main Module Structures with SCSA: (a) MobileNetv2's inverted residual module. (b) Residual blocks for ResNet-18 and ResNet-34. (c) Residual blocks for ResNet-50 and above.

Fig. 6. Visualization of semantic segmentation results using the UperNet model.

leveraging multi-semantic information to better perceive the contextual space of relevant objects.
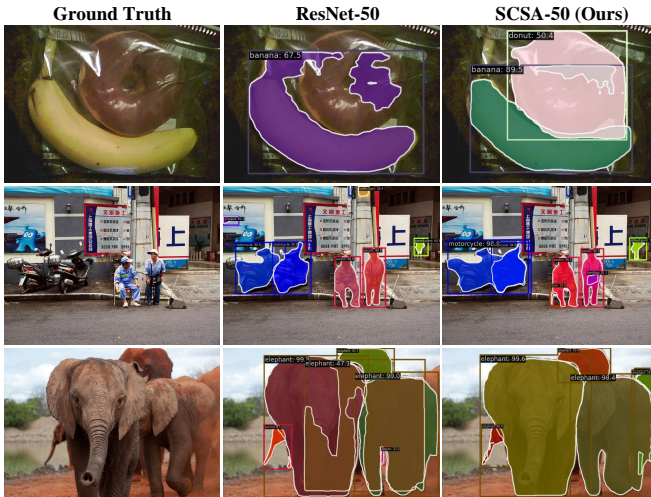

Fig. 5. Visualization of instance segmentation results using the Mask R-CNN detector. Each instance depicted in a distinct color.

B. Semantic Segmentation

In addition to instance segmentation, we visualize semantic segmentation outcomes as well. Utilizing the ADE20K dataset, we randomly select three samples for inference via UperNet and compared our method visually with the ResNet-50 baseline. It can be observed from Figure 6 that our method significantly improves the segmentation of objects that overlap and are semantically adjacent, effectively distinguishing between scenarios such as spectators seated on chairs and toilets near bathtubs.

TABLE VI
COMPARISON OF OUR SCSA, BASED ON RESNET-50 AND RESNET-101, WITH OTHER ATTENTION MECHANISMS FOR OBJECT DETECTION PERFORMANCE ACROSS FOUR DIFFERENT DATASETS.

Datasets	Methods	ResNet-50			ResNet-101		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
VOC 0712	-	50.7	81.9	55.7	54.3	83.8	61.0
	+SE	50.2	81.9	54.1	53.7	83.6	60.1
	+ECA	50.7	82.2	55.0	54.4	84.4	60.5
	+FCA	50.8	82.0	55.2	53.7	83.7	59.6
	+CA	51.8	82.5	56.5	55.4	84.2	61.5
	+SCSA	**53.0**	**83.0**	**58.0**	**55.5**	**84.6**	**61.8**
VisDrone2019	-	22.1	37.3	23.1	23.1	38.5	24.5
	+SE	21.6	36.7	22.4	22.1	37.6	23.1
	+FCA	21.9	37.1	22.7	22.4	38.0	22.8
	+ECA	21.9	37.3	22.7	22.6	38.3	22.9
	+CA	22.8	38.3	23.9	**23.5**	39.2	**24.4**
	+SCSA	**22.9**	**38.7**	**24.0**	23.3	**39.2**	24.2
ExDark	-	39.2	71.4	38.6	42.4	74.9	43.4
	+ECA	37.9	70.7	37.2	42.4	75.1	42.8
	+SE	38.3	71.1	37.1	41.8	74.8	42.0
	+FCA	38.3	71.4	37.6	41.9	75.0	42.4
	+CA	39.5	72.2	39.8	**43.2**	75.6	**45.4**
	+SCSA	**40.2**	**73.2**	**40.0**	43.0	**75.6**	44.9
FLIR-ADAS2	-	24.7	42.2	25.5	**26.3**	**44.6**	**28.0**
	+CA	24.2	42.2	25.0	25.5	43.7	26.8
	+FCA	24.4	41.5	25.8	24.7	42.0	25.9
	+SE	24.5	**42.5**	25.5	25.2	42.9	26.0
	+ECA	24.6	41.9	25.6	25.3	42.8	25.9
	+SCSA	**24.8**	42.3	**26.1**	25.4	43.2	26.2

detector, we randomly select two samples from MSCOCO val2017 and conduct comparisons with the ResNet-50 baseline. As shown in Figure 7, our method demonstrates superior performance in challenging scenarios, including obstruction, dense environments, clusters of small objects, and low-light conditions.

APPENDIX C
VISUALIZATION OF DETECTION RESULTS

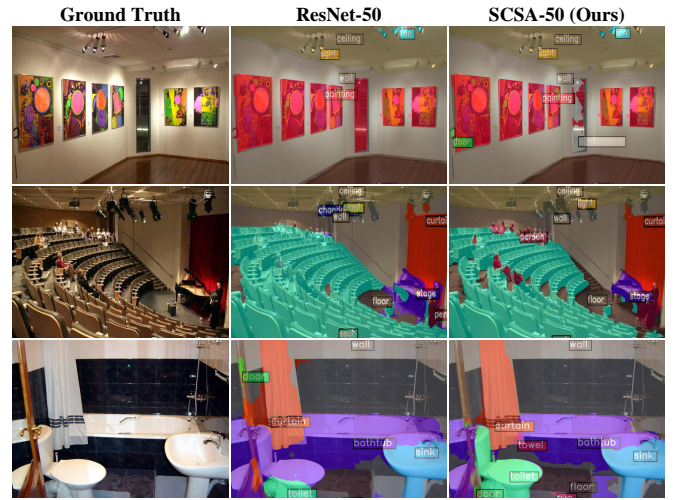We visualize the detection results using detectors such as Faster R-CNN, Cascade R-CNN, and RetinaNet. For each

APPENDIX D
MORE EXPERIMENTS

We are keen to explore whether attention mechanisms can be more effectively applied to various complex scene tasks.
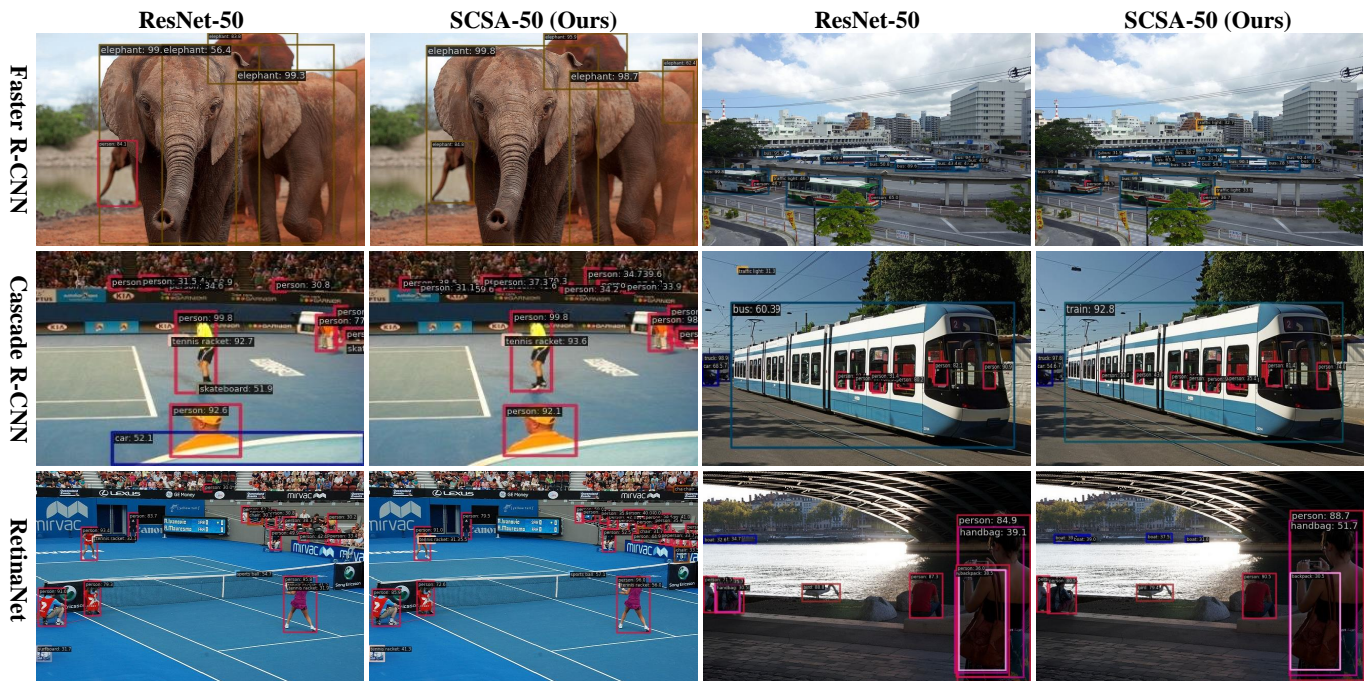
Fig. 7. Detection results are visualized on Faster R-CNN, Cascade R-CNN, and RetinaNet by respectively selecting two random samples from COCO val2017 and comparing our SCSA with the ResNet-50 baseline to demonstrate the effectiveness of our method.

TABLE VII
CODE AND DATASETS ASSETS USED IN OUR WORK.

Name	URL
ImageNet-1K	https://image-net.org/
MSCOCO 2017	https://cocodataset.org/
ADE20K	http://sceneparsing.csail.mit.edu/
VOC 0712	http://host.robots.ox.ac.uk/pascal/VOC/
VisDrone2019-DET	https://github.com/VisDrone/VisDrone-Dataset
ExDark	https://github.com/cs-chan/Exclusively-Dark-Image-Dataset
FLIR-ADAS2	https://www.flir.com/oem/adas/adas-dataset-form/
MMPretrain	https://github.com/open-mmlab/mmpretrain
MMDetection	https://github.com/open-mmlab/mmdetection
MMSegmentation	https://github.com/open-mmlab/mmsegmentation

While previous research has shown good performance on general large-scale datasets, the effectiveness in dense, low-light, and small-object scenes remains uncharted. Therefore, we conducted more experiments using representative datasets: the small-object dataset **VisDrone2019** [48], low-light dataset **ExDark** [49], infrared automotive dataset **FLIR-ADAS2** [50], and general dataset **VOC 0712** [51]. The experimental setup was consistent with the detection configurations in Section IV. As shown in Table VI, it is gratifying that our SCSA outperformed others across these datasets, further demonstrating the robustness of our strategy in maintaining channel counts and the synergistic concept of multi-semantic information. Furthermore, our results indicate that the application of attention mechanisms on long-tail datasets, such as FLIR-ADAS2, has led to minimal performance gains and even declines. This may be due to the attention mechanism's squeeze-and-excitation strategy being ill-suited for handling imbalanced distributed data, resulting in a focus on high-frequency categories while neglecting the learning of low-frequency ones. We believe that

our approach will also perform well in other detection and segmentation tasks.

APPENDIX E
FURTHER ANALYSIS OF EFFECTIVENESS

Due to inherent semantic disparities across various objects and scales, our approach leverages these multi-semantic disparities. We employ shareable depth-wise convolutions to independently learn unique patterns in different features, effectively capturing the distinctiveness of each object. Additionally, by utilizing spatial knowledge that can perceive fine-grained and coarse-grained features of different objects, our method guides strong feature interactions and recalibration. This process integrates and alleviates the semantic ambiguities caused by semantic disparities, motivating our synergistic idea. For objects of different categories or the same category at different scales, the SMSA module extracts multi-semantic spatial knowledge, while the PCSA module efficiently integrates multi-scale semantic information. This
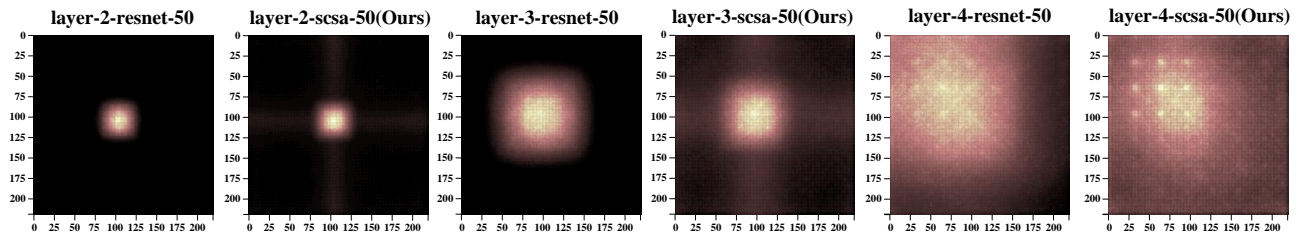
Fig. 8. Comparison of effective receptive fields(ERFs). Our SCSA provides a larger effective receptive field compared to the baseline, and the effect becomes more pronounced as the layers deepen.

enriches the global context, providing more comprehensive support for local decisions. Precise handling of local details helps selectively emphasize pixel-level areas of interest, which is a key factor in our significant advantages in object detection, instance segmentation, and semantic segmentation tasks.

A. Visualization of ERFs

As depicted in Figure 8, leveraging the spatial structure of multi-semantic modeling, our SCSA has achieved a broader perceptual area. A larger effective receptive field(ERF) is beneficial for the network to utilize rich contextual information for collective decision-making, which is one of the important factors for performance improvement. To verify that the performance of our method benefits from a larger ERF, we randomly sample 300 images of different categories from the ImageNet-1K validation set, measure the contribution of each pixel on the original image to the center point of the output feature maps of the third and fourth stages of the model, and quantify the range of the ERF with the gradient values weighted and normalized. The visualization results demonstrate that as the network layers deepen, the ERF of our SCSA becomes increasingly evident, confirming our hypothesis and the effectiveness of our method.

B. Computational Complexity

Given an input $X \in \mathbb{R}^{C \times H \times W}$, a pooling size of $P \times P$, and a depth-wise convolutional kernel size of $K \times K$, we sequentially consider the impact of dimension decoupling, depth-shared 1D convolutions, normalization, progressive compression, and channel-wise self-attention, which collectively constitute the SCSA module. For simplicity of observation, we ignore the coefficients. The computational complexities of SCSA are:

$$
\begin{aligned}
\Omega(SCSA) = {} & \mathcal{O}(HC + WC) + \mathcal{O}(KHC + KWC) \\
& + \mathcal{O}(HWC) + \mathcal{O}(P^2 H^{'} W^{'} C + H^{'} W^{'} C) \\
& + \mathcal{O}(H^{'} W^{'} C + H^{'} W^{'} C^2)
\end{aligned}
\tag{14}
$$

$H^{'}$ and $W^{'}$ denote the height and width, respectively, of the intermediate feature map produced by the progressive compression operation.

We observe that when the model width (i.e., the number of channels, C) is moderate, $\Omega(SCSA)$ scales linearly with the length of the input sequence. This indicates that our SCSA can perform inference with linear complexity when the model width is moderate.

C. Inference Throughput Evaluation

As demonstrated in Tables III and IV, we evaluate the throughput of SCSA's individual components in ablation experiments and compare the throughput across various benchmark models using different attention mechanisms. We obtain results using an GeForce RTX 4090 GPU at 224x224 resolution. Specifically, As illustrated in Table III, although SCSA is slightly slower than pure channel attention, it outperforms most hybrid attention mechanisms, including CBAM, SANet, EMA, and CPCA, and achieves the highest accuracy. Table III indicates our design optimizes balance in model complexity, inference speed, and accuracy.

APPENDIX F
LIMITATIONS

We demonstrated that our SCSA, a plug-and-play synergistic attention method, excels in image classification, object detection, and instance and semantic segmentation tasks. Although we are committed to exploring the synergistic effects across various dimensions and have empirically validated the effectiveness of leveraging multi-semantic spatial information to guide channel recalibration and enhance feature interactions for mitigating semantic differences, inference latency remains a significant challenge in real-world deployment. Our approach achieves an optimal balance of model parameters, accuracy, and inference speed at an appropriate model width. However, at larger widths, the primary bottleneck in inference speed is the use of depth-wise convolutions within the construction of a mutli-semantic spatial structure, which have low FLOPS, frequently access memory, and exhibit low computational density [26]. We believe that the positioning and quantity of attention modules should be optimized based on specific tasks and scenarios to ensure peak performance. In the future, we will explore attention synergy across multiple dimensions to ensure that attention mechanisms across different dimensions complement and enhance each other.