# Efficient Multi-Scale Attention Module with Cross-Spatial Learning

# 高效多尺度注意力模块与跨空间学习

Daliang Ouyang, Su He, Jian Zhan, Huaiyong Guo,

欧阳大力, 何苏, 占剑, 郭怀勇,

Zhijie Huang, Mingzhu Luo, Guozhong Zhang

黄志杰, 罗明珠, 张国忠

AEROSPACE SCIENCE & INDUSTRY SHENZHEN (GROUP) CO., LTD.,

航空航天科学与工业深圳（集团）有限公司，

Shenzhen, China

深圳，中国

# Abstract

# 摘要

Remarkable effectiveness of the channel or spatial attention mechanisms for producing more discernible feature representation are illustrated in various computer vision tasks. However, modeling the cross-channel relationships with channel dimensionality reduction may bring side effect in extracting deep visual representations. In this paper, a novel efficient multi-scale attention (EMA) module is proposed. Focusing on retaining the information on per channel and decreasing the computational overhead, we reshape the partly channels into the batch dimensions and group the channel dimensions into multiple sub-features which make the spatial semantic features well-distributed inside each feature group. Specifically, apart from encoding the global information to re-calibrate the channel-wise weight in each parallel branch, the output features of the two parallel branches are further aggregated by a cross-dimension interaction for capturing pixel-level pairwise relationship. We conduct extensive ablation studies and experiments on image classification and object detection tasks with popular benchmarks (e.g., CIFAR-100, ImageNet-1k, MS COCO and VisDrone2019) for evaluating its performance.

在各种计算机视觉任务中，通道或空间注意力机制在生成更可辨识的特征表示方面表现出显著的有效性。然而，使用通道维度降低来建模跨通道关系可能在提取深度视觉表示时带来副作用。在本文中，我们提出了一个新颖的高效多尺度注意力（EMA）模块。专注于保留每个通道上的信息并减少计算开销，我们将部分通道重塑为批次维度，并将通道维度分组为多个子特征，这使空间语义特征在每组特征内部均匀分布。具体来说，除了在每个并行分支中编码全局信息以重新校准通道权重外，两个并行分支的输出特征还通过跨维度交互进一步聚合，以捕捉像素级的成对关系。我们在图像分类和目标检测任务上进行了广泛的消融研究和实验，使用了流行的基准（例如，CIFAR-100、ImageNet-1k、MS COCO和VisDrone2019）来评估其性能。

# 1. Introduction

# 1. 引言

Following the evolution of deep Convolutional Neural Networks (CNNs), more notable network topologies are employed in the fields of image classification and object detection tasks. It behaves the remarkable ability to enhance the learnt feature representation when we extend the CNNs to across multiple convolutional layers. However, it leads to stack more deep convolutional counterparts and needs much consumption of memory and computation resources, which is the primary drawback for constructing the deep CNNs [1], [2]. As an alternative way, the attention mechanism method, due to the flexible structure characteristics, not only strengths the learning of more discriminative feature representation, but also can be easily plugged into backbone architecture of the CNNs. Consequently, the attention mechanisms attract much interest in the research communities of computer vision.

随着深度卷积神经网络（CNNs）的发展，更显著的网络拓扑结构被应用于图像分类和目标检测任务领域。当我们将CNNs扩展到多个卷积层时，它表现出增强学习特征表示的显著能力。然而，这导致了更多深度卷积层的堆叠，并需要大量的内存和计算资源消耗，这是构建深度CNNs的主要缺点[1]，[2]。作为一种替代方法，注意力机制方法由于其灵活的结构特性，不仅增强了对更具区分性的特征表示的学习，而且可以轻松插入到CNNs的主干架构中。因此，注意力机制在计算机视觉研究社区中引起了极大的兴趣。

It has been generally accepted that there are mainly three types of attention mechanisms proposed like the channel

人们普遍认为，主要提出了三种类型的注意力机制，如通道注意力、空间注意力及其结合。

attention, the spatial attention and both of them. As the representative channel attention, Squeeze-and-excitation (SE) explicitly modeled the cross-dimension interaction for extracting the channel-wise attention [3]. Convolutional block attention module (CBAM) [4] established the cross-channel and cross-spatial information with the semantic inter-dependencies between spatial and channel dimensions in the feature maps. Consequently, CBAM shown great potential in integrating cross-dimensional attention weights into the input features. However, the manual design of the pooling operations involves complex processing that brings in some computational overhead. To overcome the shortcomings of computational cost limitations, a long-standing and effective way, using feature grouping method to divide features into multi-group on different resources, is provided [5]. Obviously, it can make each set of features well-distributed over the space. Following the setting, Spatial group-wise enhance (SGE) attention [6] grouped the channel dimensions into multiple sub-features, and improved the spatial distribution of different semantic sub-features representations, which achieves outstanding performance.

作为代表性的通道注意力，Squeeze-and-excitation（SE）明确地建模了跨维度交互以提取通道注意力[3]。卷积块注意力模块（CBAM）[4]通过特征图中空间和通道维度之间的语义相互依赖性建立了跨通道和跨空间信息。因此，CBAM在将跨维度注意力权重整合到输入特征中展示了巨大的潜力。然而，池化操作的手动设计涉及复杂的处理，带来了一些计算开销。为了克服计算成本限制的缺点，提供了一种长期有效的方法，即使用特征分组方法将特征分成不同资源上的多组[5]。显然，它可以使每组特征在空间上分布良好。按照这一设置，空间组增强（SGE）注意力[6]将通道维度分成多个子特征，并改善了不同语义子特征表示的空间分布，从而实现了卓越的性能。

One of the most effective ways to manage model complexity is to use the convolution with channel dimensionality reduction [7]. Comparing with the SE attention, Coordinate attention (CA) [8] embedded the direction-specific information into the channel attention along spatial dimension direction, and selected an appropriate reduction ratio of channel dimensionality achieving comparable performance. On the contrary, such phenomenon is probably the most common problem that alleviates the computational burden with dimensionality reduction in the pixel-wise regression as compared with the coarse-grained CV tasks. Inspired by the thoughts that estimate the highly non-linear pixel-wise semantics, the Polarized self-attention (PSA) [9] completely collapsed the input feature maps along the counterpart channel dimensions and reserved high spectral resolution. With a small reduction ratio, PSA shown great potential in performance improvement. Although the appropriate channel reduction ratios yield better performance, it may bring side effect in extracting deep visual representations, which is explored the efficiency without dimensionality reduction in Efficient channel attention (ECA) [10].

管理模型复杂性的最有效方法之一是使用带有通道维度减少的卷积 [7]。与SE注意力机制相比，坐标注意力（CA）[8] 将方向特定信息嵌入到沿空间维度方向的通道注意力中，并选择了一个适当的通道维度减少比例，以达到可比较的性能。相反，这种现象可能是减轻像素级回归计算负担的最常见问题，与粗粒度的计算机视觉任务相比，这种回归通过维度减少来实现。受到估计高度非线性的像素级语义的思想启发，极化自注意力（PSA）[9] 完全沿对应通道维度折叠输入特征图，并保留了高光谱分辨率。PSA 以较小的减少比例，在性能提升上显示出巨大的潜力。尽管适当的通道减少比例能够带来更好的性能，但它可能在提取深度视觉表示时带来副作用，这在高效的通道注意力（ECA）[10] 中探索了在不进行维度减少的情况下提高效率的方法。

Large layers depth plays an important role in increasing the representational ability of the CNNs. However, it inevitably leads to more sequential processing and higher latency [11], [12]. Different from the large depth attentions described as a linear sequence, Triplet attention (TA) [13] proposed a triplet parallel branches structure for capturing cross-dimension interaction against the different parallel branches. With the parallel substructures, Shuffle attention (SA) [14] grouped channel dimensions into multiple sub-features and addressed them in parallel, which can be efficiently parallelized across multiple processors. Moreover, Parallel networks (ParNet) [15] constructed the parallel sub-networks improving the efficiency of feature extraction while maintaining small depth and low latency.

大层深度在提高卷积神经网络（CNNs）表示能力方面发挥着重要作用。然而，这不可避免地导致更多的顺序处理和更高的延迟 [11]，[12]。与描述为线性序列的大深度注意力不同，三元组注意力（TA）[13] 提出了一个三元组并行分支结构，用于捕获不同并行分支间的跨维度交互。利用并行子结构，洗牌注意力（SA）[14] 将通道维度分组为多个子特征，并并行处理它们，这可以在多个处理器上有效地并行化。此外，并行网络（ParNet）[15] 构建了并行子网络，提高了特征提取的效率，同时保持了小层深度和低延迟。

Taking the inspiration from the aforementioned attention mechanisms, it can be seen that the cross-dimensional interaction contributes to the channel or spatial attention prediction. We, based on the grouping structure, revise the sequential processing method of CA and propose a novel efficient multi-scale attention (EMA) without dimensionality reduction. Note that here, only two convolutional kernels will be placed in the parallel subnetworks respectively. One of the parallel subnetworks is a convolutional kernel that handles in the same manner as shown in and the other is a convolutional kernel. To demonstrate the generality of our proposed EMA, the detailed experiments are presented in Section 4, including the results on the CIFAR-100, ImageNet-1k, COCO and VisDrone2019 benchmarks. Together with the experiment results on image classification and object detect tasks are shown in Figure. 1. Our main contributions are concluded as follows:

从上述注意力机制的启示中，可以看出跨维度交互有助于通道或空间注意力预测。基于分组结构，我们对CA的顺序处理方法进行了修订，并提出了一种无需降维的新型高效多尺度注意力（EMA）。请注意，在这里，将在并行子网络中分别放置两个卷积核。其中一个并行子网络是一个 卷积核，其处理方式与 中所示相同，另一个是 卷积核。为了证明我们提出的EMA的通用性，第4节展示了详细的实验，包括在CIFAR-100、ImageNet-1k、COCO和VisDrone2019基准上的结果。与图像分类和目标检测任务的实验结果一起显示在图1中。我们的主要贡献可以概括如下：

* We propose a novel cross-spatial learning method and design a multi-scale parallel subnetworks for establishing both short and long-range dependencies.
* 我们提出了一种新的跨空间学习方法，并设计了一个多尺度并行子网络，用于建立短范围和长范围的依赖关系。
* We consider a generic method that reshapes the partly channel dimensions into the batch dimensions to avoid some form of dimensionality reduction via a universal convolution.
* 我们考虑了一种通用方法，通过通用卷积将部分通道维度重塑为批次维度，以避免某种形式的维度降低。
* Apart from building the local cross-channel interaction in each parallel subnetwork without channel dimensionality reduction, we also fuse the output feature maps of the two parallel subnetworks by a cross-spatial learning method.
* 除了在每个并行子网络中构建局部跨通道交互而不降低通道维度外，我们还通过跨空间学习方法融合了两个并行子网络的输出特征图。
* Comparing with CBAM, Normalization-based Attention Module (NAM) [16], SA, ECA and CA, EMA not only achieves the better results, but also is more efficient in terms of required parameters.
* 与CBAM、基于归一化的注意力模块（NAM）[16]、SA、ECA和CA相比，EMA不仅实现了更好的结果，而且在所需参数方面更加高效。

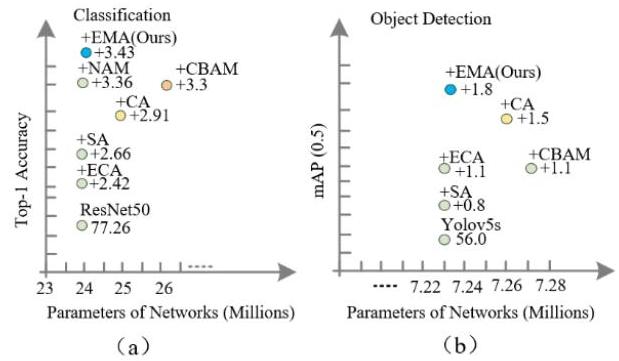


Fig.1. Comparing the accuracy of different attention methods with ResNet50 [17] as backbones, it shows EMA behaves the highest Top-1 accuracy while having less model complexity. Models are also evaluated on the COCO benchmark with the standard backbone of yolov5s (v6.0) [18], which illustrates that EMA is efficient yet effective.

图1. 比较了不同注意力方法在ResNet50 [17] 作为基础网络时的准确性，结果显示EMA在保持较低模型复杂度的同时，具有最高的Top-1准确性。模型也在使用yolov5s（v6.0）[18] 作为标准基础网络的COCO基准上进行了评估，这表明EMA既高效又有效。

# 2. Related Work

# 2. 相关工作

Feature grouping. Feature grouping has been studied extensively in the previous literature. To alleviate the restriction of computer computing budgets, AlexNet illustrated the grouped convolution is favorable to distribute the model over two groups on more GPU resources [19]. With increasing the number of feature grouping, ResNeXt has been proved that the activation of the sub-features will be spatially affected by different patterns and noisy backgrounds [20]. Res2Net was assumed as a hierarchical mode to transfer the grouped sub-features enabling CNNs for representing features at multiple scales [21]. Focusing on the feature grouping structure, SGE exploited the overall information of the entire group space to both strengthen the feature learning in semantic regions and compress the noise, but it failed in modeling the correlation between spatial and channel attention. To emphasize meaningful representation power of considering both channel and spatial attention features, SA divided the channel dimensions into multiple groups, and introduced those groups into two parallel branches with an equalitarian distribution method. Hence, the two branches can model the correlation between spatial and channel attention information separately. However, only part of the channels will be taken into account to exploit the inter-relationship of channels and construct informative features by fusing both spatial and channel-wise information.

特征分组。特征分组在之前的文献中已被广泛研究。为了缓解计算机计算预算的限制，AlexNet说明了分组卷积有利于在更多的GPU资源上分布模型至两个组 [19]。随着特征分组数量的增加，ResNeXt已经证明子特征的激活将受到不同模式和噪声背景的空间影响 [20]。Res2Net被假设为一种分层模式，用以转移分组子特征，使CNN能够表示多尺度的特征 [21]。专注于特征分组结构，SGE利用了整个组空间的总体信息，以加强语义区域内的特征学习并压缩噪声，但它未能建模空间和通道注意力之间的相关性。为了强调同时考虑通道和空间注意力特征的有意义表示能力，SA将通道维度分为多个组，并将这些组引入两个并行分支，采用等分配方法。因此，这两个分支可以分别建模空间和通道注意力信息之间的相关性。然而，只有部分通道会被考虑来利用通道间的关系，并通过融合空间和通道信息来构建信息特征。

Multi-stream networks. Given the sense that stacking one layer after another increase depth of the network for learning increasingly abstract features. As a fine-grained attention mechanism, the parallel structure was utilized by PSA to model the long-range dependencies towards high-quality pixel-wise regression task and yielded remarkable gains. Although the parallel substructures strengthen the capacity of visual representation, they bring a number of additional parameters and calculations, which is less suitable for applications. Correspondingly, the triplet attention blended cross-channel and spatial information with rotation operation into the three parallel branches for learning increasingly abstract features. However, the captured attention weights are directly aggregated by simple averaging, which is unfavorable to boost the discriminability of deep features.

多流网络。考虑到层层堆叠可以增加网络的深度，以学习越来越抽象的特征。作为一种细粒度的注意力机制，PSA利用并行结构来模拟对高质量像素级回归任务的长距离依赖，并取得了显著的效果。尽管并行子结构增强了视觉表征的能力，但它们带来了大量的额外参数和计算，这在应用中不太合适。相应地，三元组注意力将跨通道和空间信息通过旋转操作融合到三个并行分支中，以学习越来越抽象的特征。然而，捕获的注意力权重直接通过简单平均进行聚合，这对提高深度特征的可辨识性是不利的。

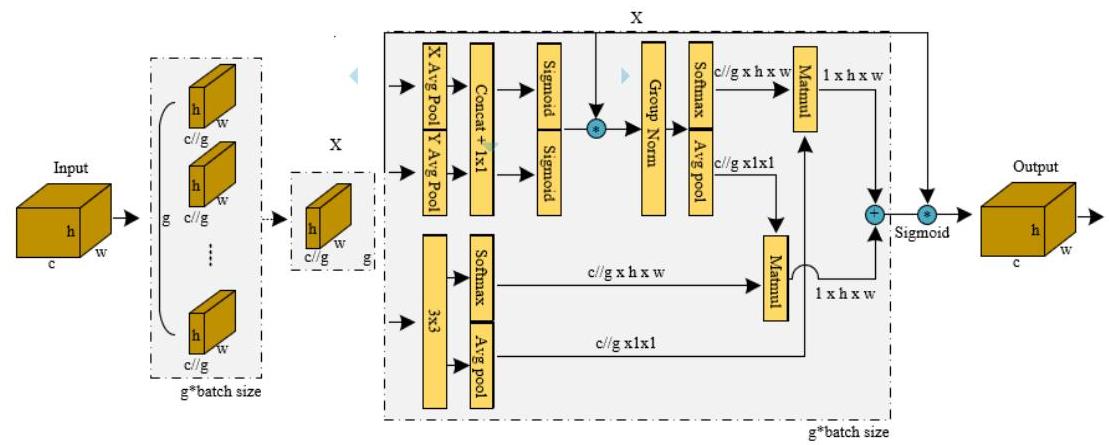


Fig.2. Illustration of our proposed EMA. Here, "g" means the divided groups, "X Avg Pool" represents the 1D horizontal global pooling and "Y Avg Pool" indicates the 1D vertical global pooling, respectively.

图2. 我们提出的EMA的说明。在这里，“g”表示分组的组数，“X Avg Pool”代表一维水平全局池化，“Y Avg Pool”表示一维垂直全局池化。

Multi-scale convolution. Different kernel sizes can enable the CNNs to collect multi-scale spatial information in the same processing stage. To enrich the feature space, Inception [22] presented multi-branch structure, where the local receptive fields in each branch are not fixed. Thus, the aggregation approach enables the CNNs to aggregate multi-scale information from different branches. Selective kernel networks [23] adopted an adaptive selection strategy that realizes adaptive receptive field size of neurons to effectively enriching feature representations. In addition, EPSANet [24] replaced the convolution with the established multi-scale pyramid structure, which models a cross-channel interaction in a local manner and learns the multi-scale spatial information independently.

多尺度卷积。不同的卷积核大小可以使CNN在同一处理阶段收集多尺度空间信息。为了丰富特征空间，Inception [22] 提出了多分支结构，其中每个分支的局部感受野不是固定的。因此，聚合方法使得CNN能够从不同的分支中聚合多尺度信息。选择性核网络 [23] 采用了自适应选择策略，实现了神经元自适应感受野大小，以有效丰富特征表示。此外，EPSANet [24] 将 卷积替换为建立的多尺度金字塔结构，该结构以局部方式模拟跨通道交互，并独立地学习多尺度空间信息。

Comparing with the aforementioned attention modules, our proposed multi-scale attention module shows more better performance improvements. Different to the above attention methods, where the learnt attention weights are aggregated by a simple averaging method, we fuse the learnt attention maps of the parallel subnetworks by a cross-spatial learning method. It uses the matrix dot-product operations aiming at capturing pixel-level pairwise relationship and highlighting global context for all pixels [25], [26].

与上述注意力模块相比，我们提出的多元尺度注意力模块显示出更好的性能提升。不同于上述注意力方法，其中学习到的注意力权重是通过简单的平均方法聚合的，我们通过交叉空间学习方法融合了并行子网络的注意力图。它使用矩阵点积操作，旨在捕捉像素级的成对关系并为所有像素突出全局上下文 [25]，[26]。

# 3. Efficient Multi-Scale Attention

# 3. 高效多元尺度注意力

In this section, we first revisit the coordinate attention block, where the positional information is embedded into the channel attention maps for blending cross-channel and spatial information. We will develop and analyze our proposed EMA module, in which the parallel subnetworks block helps effectively capture the cross-dimension interaction and establish the inter-dimensional dependencies.

在这一节中，我们首先重新审视坐标注意力模块，其中位置信息被嵌入到通道注意力图中，以融合跨通道和空间信息。我们将开发和分析我们提出的EMA模块，在该模块中，并行子网络块有助于有效地捕捉跨维度交互并建立维度间的依赖关系。

# 3.1. Revisit Coordinate Attention (CA)

# 3.1. 重新审视坐标注意力（CA）

As shown in Figure. 3 (a), CA block can be firstly viewed as a similar approach to the SE attention module, where the global average-pooling operation is exploited to model the cross-channel information. Generally, the channel-wise statistics can be generated by using a global average pooling, where the global spatial position information is squeezed into a channel descriptor. Subtly different to the SE, CA embedded the spatial positional information into channel attention maps for the enhancement of feature aggregation.

如图3（a）所示，CA模块首先可以看作是与SE注意力模块类似的方法，其中利用全局平均池化操作来建模跨通道信息。通常，可以通过使用全局平均池化生成通道-wise统计量，其中全局空间位置信息被压缩成一个通道描述符。与SE微妙的区别在于，CA将空间位置信息嵌入到通道注意力图中，以增强特征聚合。

Note that the CA will decomposes the original input tensors into two parallel 1D feature encoding vectors for modeling the cross-channel dependencies with spatial positional information. Firstly, one of the parallel routes is directly from a 1D global average-pooling along the horizontal dimension direction and hence can be viewed as a collection of positional information along the vertical dimension direction [8]. Let the original input tensor denotes the intermediate feature map, where means the numbers of the input channels, and indicate the spatial dimensions of the input features respectively. Consequently, the 1D global average-pooling for encoding the global information along the horizontal dimension direction in at height can be denoted by

注意，CA会将原始输入张量分解为两个并行的1D特征编码向量，用于建模带有空间位置信息的跨通道依赖。首先，其中一个并行路径直接来自沿水平维度方向进行的1D全局平均池化，因此可以看作是沿垂直维度方向的位置信息的集合[8]。设原始输入张量 表示中间特征图，其中 表示输入通道的数量， 和 分别表示输入特征的空间维度。因此，在高度 处沿水平维度方向编码全局信息的1D全局平均池化可以表示为

where indicates the input features at -th channel. With such encoding processes, CA captures the long-range dependencies at the horizontal dimension direction and preserves precise positional information at the vertical dimension direction. Similarly, the other one route is directly from a 1D global average-pooling along the horizontal dimension direction and hence can be viewed as a collection of positional information along the vertical dimension direction. The route utilizes the 1D global average-pooling along the vertical dimension direction to capture long-range interactions spatially and preserve the precise positional information along the horizontal dimension direction, strengthening the attention to the spatial region of interest. The pooling output in at width can be formulated as其中 表示 通道的输入特征。通过这样的编码过程，CA捕获了水平维度方向上的长距离依赖，并在垂直维度方向上保留了精确的位置信息。同样，另一个路径直接来自沿水平维度方向的1D全局平均池化，因此也可以看作是沿垂直维度方向的位置信息的集合。该路径利用沿垂直维度方向的1D全局平均池化来捕获空间上的长距离交互，并保留沿水平维度方向上的精确位置信息，增强了对感兴趣空间区域的关注。在宽度 处的池化输出可以表示为

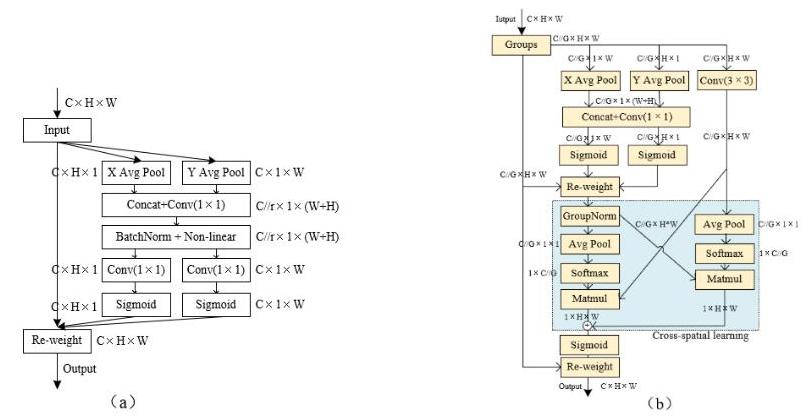


Fig.3. Comparisons with different attention modules: (a) CA Module; (b) EMA module.

图3. 与不同注意力模块的比较：(a) CA模块；(b) EMA模块。

where indicates the input features at -th channel. In the following, the input features can encode the global feature information and assist the model in capturing global information along two spatial directions respectively, which are in the absent of convolutions. Moreover, it generates two parallel 1D feature encoding vectors, and then permutes one vector into the other vector shape before concatenating two parallel 1D feature encoding vectors across a convolutional layer. Those two parallel 1D feature encoding vectors will share a convolutional convolution with dimensionality reduction. The kernel is to enable model to capture local cross-channel interaction and share the similarities with channel-wise convolutions. And then, CA further factorizes the outputs of convolution kernel into two parallel 1D feature encoding vectors and stack one convolutional convolution followed by a non-linearity Sigmoid function in each parallel routes respectively. Finally, the learnt attention map weights of the two parallel routes will be utilized to aggregate the raw intermediate feature map as the final outputs. Therefore, CA not only preserved the precise positional information, but also effectively exploit the long-range dependencies by encoding inter-channel and spatial information.

其中 表示第 个通道的输入特征。以下内容中，输入特征可以编码全局特征信息，并帮助模型分别在两个空间方向上捕捉全局信息，而不使用卷积。此外，它生成了两个并行的1D特征编码向量，然后在跨卷积层连接两个并行的1D特征编码向量之前，将一个向量排列为另一个向量的形状。这两个并行的1D特征编码向量将共享一个 维度降低的卷积。该 核使模型能够捕捉局部跨通道交互并共享通道卷积的相似性。然后，CA进一步将 卷积核的输出分解为两个并行的1D特征编码向量，并在每个并行路径中分别堆叠一个 卷积和一个非线性Sigmoid函数。最终，两个并行路径学到的注意力图权重将用于汇总原始中间特征图作为最终输出。因此，CA不仅保留了精确的位置信息，而且通过编码通道间和空间信息有效地利用了长距离依赖性。

It is self-evident that the CA embeds the precise positional information into channel-wise and captures long-range interactions spatially, achieving the impressive performance. The two 1D global average-pooling are designed for encoding the global information along two spatial dimensions direction and capture the long-range interactions spatially along different dimension directions respectively. However, it neglects the importance of the interaction among entirely spatial positions. Moreover, the limited receptive field of kernel convolution adversely hinders for modeling of local cross-channel interaction and capitalizing on the contextual information.

显然，CA将精确的位置信息嵌入到通道中，并在空间上捕捉长距离交互，实现了令人印象深刻的表现。两个1D全局平均池化是为了编码两个空间维度方向的全局信息并分别沿着不同维度的方向在空间上捕捉长距离交互。然而，它忽视了完全空间位置之间交互的重要性。此外， 核卷积有限的感受野对建模局部跨通道交互和利用上下文信息产生了不利影响。

# 3.2. Multi-Scale Attention (EMA) Module

# 3.2. 多尺度注意力（EMA）模块

The parallel substructures help the networks avoid more sequential processing and large depth. Given the above defined parallel processing strategy, we adopt it in our EMA module. The overall structure of EMA is shown in Figure. 3 (b). In this section, we will discuss how the EMA learns effective channel descriptions without channel dimensionality reduction in convolution operations, and produce a better pixel-level attention for high-level feature maps. Specifically, we only pick out the shared component of convolution from the CA module, named it as branch in our EMA. To aggregate multi-scale spatial structure information, a 3x3 kernel is placed in parallel with branch for fast responses and we name it as branch. Considering the feature grouping and multi-scale structures, it is favorable to efficiently establish both short and long-range dependency for better performance.

并行子结构帮助网络避免更顺序的处理和大的深度。考虑到上述定义的并行处理策略，我们在EMA模块中采用它。EMA的整体结构如图3（b）所示。在本节中，我们将讨论EMA如何在不进行卷积操作中的通道维度降低的情况下学习有效的通道描述，并为高级特征图产生更好的像素级注意力。具体来说，我们只从CA模块中提取 卷积的共享部分，将其命名为EMA中的 分支。为了聚合多尺度空间结构信息，我们在 分支并行放置一个3x3的核，以获得快速响应，我们将其命名为 分支。考虑到特征分组和多尺度结构，有利于有效地建立长短距离依赖，以获得更好的性能。

Feature Grouping. For any given input feature map , EMA will divide into sub-features across the channel dimensions direction for learning different semantics, where the groups-style can be donated by . Without losing generality, we let and assumed that the learnt attention weight descriptors will be utilized to strength the feature representation of interest region in each sub-feature.

特征分组。对于任何给定的输入特征图 ，EMA将沿着通道维度方向将 分为 个子特征，以学习不同的语义，其中组风格可以表示为 。不失一般性，我们令 ，并假设学到的注意力权重描述符将被用来增强每个子特征中感兴趣区域的特征表示。

| Method | Backbone | #.Param. | FLOPs | Top-1 (%) | Top-5 (%) |
| --- | --- | --- | --- | --- | --- |
| Baseline [17] | ResNet50 |  | 1.30G | 77.26 | 93.63 |
| + CBAM [16] |  | 1.31G | 80.56 | 95.34 |
| + SA |  | 1.31G | 79.92 | 95.00 |
| ECA | 23.71M | 1.31G | 79.68 | 95.05 |
| + NAM [16] | 23.71M | 1.31G | 80.62 | 95.28 |
|  |  | 1.36G | 80.17 | 94.94 |
| + EMA (ours) | 23.85M | 1.32G | 80.69 | 95.59 |
| Baseline [17] | ResNet101 | 42.70M |  | 77.78 | 94.39 |
|  | 46.22M |  | 80.01 | 94.78 |
| + EMA (ours) | 42.96M | 2.53G | 80.86 | 95.75 |

Table 1: Comparison of various attention methods on CIFAR-100 in terms of network parameters (in millions), FLOPs, Top-1 and Top-5 Validation Accuracies(%).

表1：在CIFAR-100上，不同注意力方法在网路参数（百万）、FLOPs、Top-1和Top-5验证准确率（%）方面的比较。

Parallel Subnetworks. The large local receptive fields of neurons enable the neurons to collect multi-scale spatial information. Accordingly, EMA conducts that three parallel routes are exploited to extract attention weight descriptors of the grouped feature maps. Two of parallel routes is in branch and the third one route is that the 3x3 branch. For capturing dependencies across all channels and relieving the computation budgets, we model the cross-channel information interaction at channel direction. To be more specific, there are two 1D global average pooling operations employed to encode the channel along two spatial directions respectively in branch and only a single kernel is stacked in branch for capturing multi-scale feature representation.

并行子网络。神经元的大局部感受野使得神经元能够收集多尺度空间信息。因此，EMA实施三种并行路径来提取分组特征图的注意力权重描述符。其中两条并行路径位于 分支，第三条路径是 3x3 分支。为了捕捉所有通道之间的依赖关系并减轻计算预算，我们在通道方向上建模跨通道信息交互。具体来说， 分支中分别沿两个空间方向使用了两个一维全局平均池化操作来编码通道，而在 分支中仅堆叠了一个 核以捕捉多尺度特征表示。

Given the truth that there is no batch coefficient in the dimension of the convolution function for the normal convolution, the number of convolution kernels are independent of the batch coefficients of the forward operational inputs. For example, the parameter dimension of the normal 2D convolution kernel in Pytorch is , which is not involved the batch dimensions, where oup means the out planes of the inputs, inp indicates the input planes of the input features and denotes the kernel size respectively. Accordingly, we reshape and permute groups into the batch dimension, and redefine the input tensor with shape of . On the one hand, with similar treatment as CA, we concatenate the two encoded features against the images height direction and make it share the same convolution without dimensionality reduction in branch. After factorize the outputs of convolution into two vectors, two non-linear Sigmoid functions are employed to fit the 2D Binormial distribution upon linear convolutions. For achieving different cross-channel interactive features between the two parallel routes in branch, we aggregate the two channel-wise attention maps inside each group via a simple multiplication. On the other hand, the branch captures the local cross-channel interaction via a convolution to enlarge the feature space. In this way, EMA not only encodes the inter-channel information to adjust the importance of different channels, but also preserves the precise space structure information into channel.

鉴于正常卷积中卷积函数维度没有批次系数这一事实，卷积核的数量与前向操作输入的批次系数无关。例如，Pytorch中普通2D卷积核的参数维度为 ，这不涉及批次维度，其中oup代表输入的输出平面，inp表示输入特征的输入平面， 分别表示核大小。因此，我们将 组重塑并置换到批次维度，并重新定义输入张量的形状为 。一方面，与CA的处理方式相似，我们将两个编码特征在图像高度方向上连接，使其在 分支上共享相同的卷积而不进行维度降低。在将 卷积的输出分解为两个向量后，我们使用两个非线性Sigmoid函数来拟合线性卷积上的2D双变量分布。为了在 分支的两个并行路径之间实现不同的跨通道交互特征，我们通过简单相乘在每个组内聚合两个通道注意力图。另一方面， 分支通过 卷积捕获局部跨通道交互，以扩大特征空间。这样，EMA不仅编码了通道间的信息以调整不同通道的重要性，还保留了精确的空间结构信息到通道中。

Cross-spatial learning. Benefiting from the capability of building interdependencies among channels and spatial locations, there have been extensively studied and broadly used in a variety of computer vision tasks recently [27], [28]. In PSA, it exhausted the representation capacity within its channel-only and spatial-only branches, and kept the highest internal resolution in attention learning to solve the semantic segmentation. Inspired by this, we provide a cross-spatial information aggregation method at different spatial dimension direction for richer feature aggregation. Note that here, we still have introduced two tensors where one is the output of branch and the other is the output of the branch. Then, we utilize the global average pooling to encode global spatial information in the outputs of branch, and the outputs of the least branch will be transformed to the correspond dimension shape directly before the joint activation mechanism of channel features, i.e., [9]. The 2D global pooling operation is formulated as

跨空间学习。得益于在通道和空间位置之间建立相互依赖性的能力，近年来在多种计算机视觉任务中得到了广泛的研究和应用 [27]，[28]。在PSA中，它在仅通道和仅空间分支内耗尽了表示能力，并在注意力学习中保持最高内部分辨率来解决语义分割问题。受此启发，我们为更丰富的特征聚合提供了在不同空间维度方向上的跨空间信息聚合方法。注意，在这里，我们仍然引入了两个张量，一个是 分支的输出，另一个是 分支的输出。然后，我们使用 全局平均池化在 分支的输出中编码全局空间信息，以及最小分支的输出将在通道特征联合激活机制之前直接转换为相应的维度形状，即 [9]。二维全局池化操作可以表示为

which is designed for encoding the global information and modeling the long-range dependencies. For efficient computation, the natural non-linear functions Softmax for 2D Gaussian maps is employed at the outputs of 2D global average pooling to fit the upon linear transformations. By multiplying the outputs of above parallel processing with matrix dot-product operations, we derived our first spatial attention map. To observe this, it collects different scale spatial information in the same processing stage. Moreover,

这是为了编码全局信息并建模长距离依赖而设计的。为了高效计算，在2D全局平均池化的输出上使用了自然的非线性函数Softmax来适应上述线性变换。通过将上述并行处理的结果与矩阵点积操作相乘，我们得到了我们的第一个空间注意力图。为了观察这一点，它在同一处理阶段收集不同尺度的空间信息。此外，

| Model | Datasets | #.Param. | FLOPs | mAP (0.5) | mAP (0.5:0.95) |
| --- | --- | --- | --- | --- | --- |
| Yolov5s [18] | COCO |  |  | 56.0 | 37.2 |
| + CBAM | 7.27M |  | 57.1 | 37.7 |
| + SA |  |  | 56.8 | 37.4 |
| + ECA |  |  | 57.1 | 37.6 |
|  |  | 16.50M | 57.5 | 38.1 |
| + EMA (ours) | 7.24M | 16.53M | 57.8 | 38.4 |
| Yolov5x [30] | VisDrone | 90.96M | 314.2M | 49.29 | 30.0 |
| + CBAM | 91.31M | 315.1M | 49.40 | 30.1 |
| + CA | 91.28M | 315.2M | 49.30 | 30.1 |
| + EMA (ours) | 91.18M | 315.0M | 49.70 | 30.4 |

Table 2: Object detection results of different attention methods on COCO and VisDrone val datasets. EMA Attention results in higher performance gain with slightly higher computational overhead.

表2：不同注意力方法在COCO和VisDrone验证数据集上的对象检测结果。EMA注意力带来了更高的性能提升，但计算开销略有增加。

we similarly utilize the global average pooling to encode global spatial information in the branch and the branch will be transformed to the correspond dimension shape directly before the joint activation mechanism of channel features, i.e., . After that, the second spatial attention map, which preserves the entire precise spatial positional information is derived. Finally, the output feature map within each group is calculated as the aggregation of the two generated spatial attention weight values followed by a Sigmoid function. It captures pixel-level pairwise relationship and highlights global context for all pixels. The final output of EMA is the same size of , which is efficient yet effective to stack into modern architectures.

我们同样使用 全局平均池化来编码 分支和 分支中的全局空间信息，在通道特征联合激活机制之前， 分支将被直接转换成相应的维度形状。之后，第二个空间注意力图保留了整个精确的空间位置信息。最终，每个组内的输出特征图是两个生成的空间注意力权值值经过 Sigmoid 函数聚合的结果。它捕获像素级的成对关系并为所有像素突出显示全局上下文。EMA 的最终输出与 的大小相同，既高效又有效，可以堆叠到现代架构中。

As discussed above, we can know that the attention factors are only guided by the similarities between the global and local feature descriptors inside each group. Considering the cross-spatial information aggregation method, both the long-range dependencies will be modeled, and the precise positional information are embedded into EMA. Fusing context information with different scales enables the CNNs to produce a better pixel-level attention for high-level feature maps. Subsequently, the parallelizing of the convolution kernels seems to be a more powerful structure to handle both short and long-range dependency by using the cross-spatial learning method. In contrast to the progressive behavior of limited receptive fields formed, utilizing and convolutions in parallel capitalizes more contextual information among intermediate feature maps.

如上所述，我们可以知道注意力因素仅由每个组内全局和局部特征描述符之间的相似性引导。考虑到跨空间信息聚合方法，长距离依赖将被建模，精确的位置信息被嵌入到 EMA 中。融合不同尺度的上下文信息使 CNN 能够为高级特征图产生更好的像素级注意力。随后，卷积核的并行似乎是一种更强大的结构，通过使用跨空间学习方法处理短距离和长距离依赖。与形成的有限感受野的渐进行为相比，并行使用 和 卷积能够更好地利用中间特征图之间的上下文信息。

# 4. Experiments

# 4. 实验

In this section, we provide the details for experiments and results to demonstrate the performance and efficiency of our proposed EMA. We conduct experiments on the challenging computer vision tasks like classification on CIFAR-100 and ImageNet-1k, and object detection on MS COCO and VisDrone2019 datasets. To verify the efficient performance, we integrate EMA into the standard network architectures like ResNet50/101 and MobileNetV2 [29], respectively. Our object detect code implementation is based upon the Pytorch YOLOv5s (v6.0) repository by Ultralytics. For image classification tasks on the CIFAR- 100 and ImageNet-1k datasets, the experiments are performed with exactly the same data augmentation and training configuration settings in NAM and CA to make fair comparisons. The numbers of groups in our proposed EMA module is set as 32. All experiments run on a PC equipped with two RTX 2080Ti GPUs and on Intel(R) Xeon Silver 4112 CPU@2.60Ghz.

在本节中，我们提供了实验和结果的详细信息，以展示我们提出的EMA的性能和效率。我们在具有挑战性的计算机视觉任务上进行实验，如CIFAR-100和ImageNet-1k的数据分类，以及MS COCO和VisDrone2019数据集上的目标检测。为了验证高效的性能，我们将EMA分别集成到标准的网络架构中，如ResNet50/101和MobileNetV2 [29]。我们的目标检测代码实现基于Ultralytics的Pytorch YOLOv5s (v6.0) 存储库。在CIFAR-100和ImageNet-1k数据集上的图像分类任务中，实验使用了与NAM和CA完全相同的数据增强和训练配置设置，以进行公平比较。我们提出的EMA模块中的组数 设置为32。所有实验都在配备两块RTX 2080Ti GPU和Intel(R) Xeon Silver 4112 CPU@2.60Ghz的PC上运行。

# 4.1. Image Classification on CIFAR-100

# 4.1. CIFAR-100上的图像分类

We investigate our proposed EMA on CIFAR-100 datasets, whose sets include the images with pixels and consist of images drawn from 100 classes. The training set is comprised of images and the validation set is comprised of images. We exploit stochastic gradient descent (SGD) with momentum of 0.9 and the weight decay of . The batch size is 128 by default. Our networks of all comparing approaches are trained for 200 epochs to make fair comparisons. After the training, we evaluate the performance of our network using the standard CIFAR Top-1 and Top-5 accuracy metrics.

我们在CIFAR-100数据集上研究了我们提出的EMA，该数据集包括 像素的图像，并包含来自100个类别的图像。训练集由 图像组成，验证集由 图像组成。我们利用带有动量为0.9和权重衰减为 的随机梯度下降（SGD）。默认的批量大小为128。为了进行公平比较，我们训练所有比较方法中的网络200个周期。训练完成后，我们使用标准的CIFAR Top-1和Top-5准确度指标评估我们网络的性能。

As shown in Table 1, a comparison of several other attention mechanisms over the baseline of ResNet50/101 shows that integrating with EMA gains a very comparative performance with relatively small model complexity (i.e., network parameters and floating-point operations per second (FLOPs)). Comparing with the standard baseline of ResNet50, EMA achieves gains in terms of Top-1 accuracy and advantages over the Top-5 accuracy. With almost the same computational complexity, the Top- 1 accuracy can be improved by by our proposed EMA as compared to the CA. In addition, using ResNet101 as the backbone model, we compare EMA with CA. Obviously, our EMA outperforms CA by a large margin with less parameters and lower computational cost. It is worth noting that the gain in term of Top-1 average accuracy of CA is slightly dropped from for the ResNet50 to for the ResNet101 as the network architecture becoming deeper.

如表1所示，将EMA与ResNet50/101基线上的其他几种注意力机制进行比较，发现EMA的集成性能与相对较小的模型复杂度（即网络参数和每秒浮点运算数（FLOPs））非常接近。与ResNet50的标准基线相比，EMA在Top-1准确度上实现了 的提升，并在Top-5准确度上具有 的优势。几乎相同的计算复杂度下，与CA相比，我们提出的EMA可以将Top-1准确度提高 。此外，使用ResNet101作为基础模型，我们将EMA与CA进行了比较。显然，我们的EMA在参数更少 和计算成本更低的情况下，大幅超越了CA。值得注意的是，随着网络架构的加深，CA在Top-1平均准确度上的提升从ResNet50的 略微下降到ResNet101的 。

| Method | Backbone | #.Param. | M-Adds | Top-1 (%) | Top-5 (%) |
| --- | --- | --- | --- | --- | --- |
| Baseline [29] | MobileNet (v2) | 3.50 M | 300 M | 72.3 | 91.02 |
| + SE [8] | 3.89 M |  | 73.5 | - |
| + CBAM [8] | 3.89 M |  | 73.6 | - |
| + CA [8] | 3.95 M |  | 74.3 | - |
| + EMA (ours) | 3.55M |  | 74.32 | 91.82 |

Table 3: Comparison with attention-based models and our EMA based on MobileNetv2 on ImageNet-1k.

表3：基于注意力机制的模型和基于MobileNetv2的EMA在ImageNet-1k上的比较。

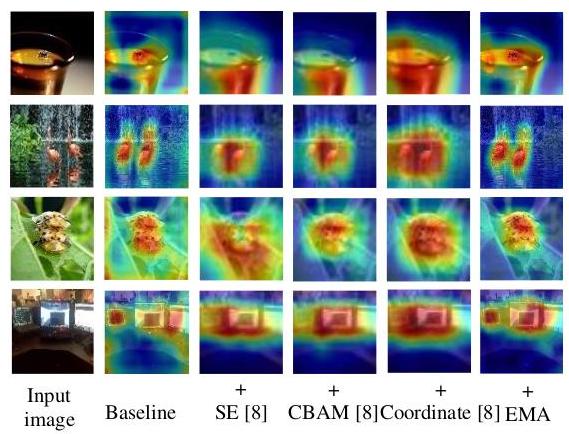


Figure 4. Visualization of feature maps produced by models with different attention-based methods in the last building block of the baseline (MobilenetV2) networks. It noted that our EMA module focuses on more relevant regions with more object details.

图4。基于不同注意力方法的模型在基线（MobilenetV2）网络的最后一个构建块中生成的特征图的可视化。值得注意的是，我们的EMA模块更加关注包含更多物体细节的相关区域。

# 4.2. Image Classification on ImageNet-1k

# 4.2. 在ImageNet-1k上的图像分类

Additionally, we compare EMA with other attention-based methods using MobileNetv2 as the baseline model on ImageNet-1k, which is a widely used large-scale benchmark for image classification. ImageNet-1K dataset consists of almost training mages and validation images with 1000 classes. During training, all images are resized to the resolution . We pick the models with best Top-1 accuracy and Top-5 accuracy performance on validation images. SGD optimizer with a learning rate of 0.4 , momentum of 0.9 , weight decay of 1e- 5 and the linear warmup learning rate of are used. Following common practice, we train MobileNetv2 for 400 epochs with batch size of 256 . To reduce stochastic noise during training, we use exponential moving average method.

此外，我们使用 MobileNetv2 作为基线模型，在 ImageNet-1k 数据集上比较了 EMA 与其他基于注意力的方法。ImageNet-1K 数据集是一个广泛使用的大型图像分类基准，包含几乎 个训练图像和 个验证图像，共 1000 个类别。在训练过程中，所有图像都被调整到 分辨率。我们选择了在验证图像上具有最佳 Top-1 准确度和 Top-5 准确度性能的模型。使用学习率为 0.4、动量为 0.9、权重衰减为 1e-5 和 的线性预热学习率的 SGD 优化器。遵循常见做法，我们以批量大小为 256 训练 MobileNetv2 达 400 个周期。为了在训练过程中减少随机噪声，我们使用了指数移动平均方法。

As shown in Table 3, EMA model outperforms all other attention-based models with comparable FLOPs or multiply-adds. The baseline of MobileNetv2 model achieves ImageNet Top-1 validation accuracy and Top-5 validation accuracy with multiply-adds. With integrating the SE module into MobileNetv2, the performance of Top-1 validation accuracy further improves to . Furthermore, the MobileNetv2 network with the CBAM module significantly improves the Top-1 validation accuracy performance over the baseline model by . Our EMA model achieves state-of-the-art performance of accuracy with multiply-adds, when compared to the baseline. The experimental results show that the params of the CA model is , while the EMA model proposed in the paper is only , which is smaller than the CA model.

如表 3 所示，EMA 模型在具有可比 FLOPs 或乘加操作的模型中优于所有其他基于注意力的模型。MobileNetv2 基线模型在 上实现了 ImageNet Top-1 验证准确度，在 上实现了 Top-5 验证准确度，具有 乘加操作。将 SE 模块集成到 MobileNetv2 中后，Top-1 验证准确度进一步提高到 。此外，带有 CBAM 模块的 MobileNetv2 网络在基线模型的基础上显著提高了 Top-1 验证准确度性能 。与基线相比，我们的 EMA 模型在 准确度上达到了最先进的性能，具有 乘加操作。实验结果表明，CA 模型的参数为 ，而本文提出的 EMA 模型仅为 ，小于 CA 模型。

# 4.3. Object Detection on MS COCO

# 4.3. 在 MS COCO 上的目标检测

To demonstrate the advantages of our proposed method in object detection, we investigate EMA on COCO datasets, whose training set is comprised of images and the apart from the fixed size of input demanded by the original settings such as resizing the images to uniform dimensions of . The epochs and the batch size are set as 300 and 50 on one RTX 2080Ti GPUs respectively. Subsequently, we provide several attention mechanisms, 0.5 to 0.95 ). To further verify the effectiveness of our method, we implement all other attention strategies into the standard Yolov5s backbone with the same default settings modules into the Yolov5s backbone both improve the performance of object detection by a clear margin. Compared to CBAM, SA and ECA models, EMA and CA gaining and our EMA performs slightly better than CA in terms of . In addition, it can be seen that the model size of EMA is , which is only lightly larger than the baseline of YOLOv5s, ECA and SA models (7.24M v.s. ). Although the FLOPs of EMA are , which are only larger than the baseline of YOLOv5s, EMA achieves the mAP (0.5) of and mAP (0.5:0.95) of on all 80 classes, which is significantly higher than other attention strategies. In general, the model size is suitable for deployment on the mobile terminals and has practical application significance.

为了展示我们提出的方法在目标检测中的优势，我们在COCO数据集上研究了EMA，其训练集包含 张图像，除了原始设置所需的固定输入尺寸，例如将图像调整到统一的尺寸 。在单个RTX 2080Ti GPU上，训练周期和批量大小分别设置为300和50。随后，我们提供了几种注意力机制（系数从0.5到0.95）。为了进一步验证我们方法的有效性，我们将所有其他注意力策略实施到标准Yolov5s骨干网络中，使用相同的默认设置。将模块集成到Yolov5s骨干网络中均明显提高了目标检测的性能。与CBAM、SA和ECA模型相比，EMA和CA性能提升明显，我们的EMA在 方面略优于CA。此外，可以看出EMA的模型大小为 ，仅比YOLOv5s的基线、ECA和SA模型的基线（7.24M v.s. ）略大。尽管EMA的FLOPs为 ，仅比YOLOv5s的基线大 ，但EMA在所有80个类别上达到了mAP (0.5) 的 和mAP (0.5:0.95) 的 ，这显著高于其他注意力策略。总的来说，该模型大小适合部署在移动终端上，并具有实际应用意义。

# 4.4. Object Detection on VisDrone

# 4.4. 在VisDrone上的目标检测

Considering our proposed EMA on the dense object detection of the multi-scale feature fusion, we add a detection head for the tiny objects based on the original YOLOv5x (v6.0) [30] and integrate EMA into prediction branch to achieve the purpose of exploring the prediction potential with self-attention mechanism. During experiments, we set the size of the input image to and we use part of pre-trained model from yolov5x for saving a lot of training time. All the attention models on VisDrone2019 trainset are trained for 300 epochs and the batch size is set as 5 . The experiments are performed with exactly the same data augmentation and training configuration settings.

考虑到我们提出的在多尺度特征融合的密集目标检测中的EMA，我们在原始的YOLOv5x（v6.0）[30]基础上为小目标增加了一个检测头，并将EMA集成到预测分支中，以实现利用自注意力机制探索预测潜力的目的。在实验中，我们将输入图像的大小设置为 ，并使用部分预训练的yolov5x模型以节省大量的训练时间。所有在VisDrone2019训练集上的注意力模型都训练了300个周期，批量大小设置为5。实验使用了完全相同的数据增强和训练配置设置。

We use the YOLOv5x as our backbone CNN for the object detection on VisDrone datasets, where the CA, CBAM and EMA attentions are integrated into the detector respectively. As observed from the results of Table 2, both the CA, CBAM and EMA can boost the baseline performance for the object detection. We can see our proposed EMA module consistently outperforms the base CA and CBAM based networks in terms of the mAP (0.5) and mAP (0.5:0.95) respectively. It is noteworthy that CBAM boosts the performance of YOLOv5x by and is higher than that of at the cost of more parameters and computations. For CA, it almost obtains the same performance as the baseline and surpass the YOLOv5x by in terms of the mAP (0.5), while CA achieves higher parameters and computations than EMA (91.28M v.s. 91.18M and v.s. ). Specifically, EMA adds more parameters than baseline method, which have an improvement of over YOLOv5x on mAP (0.5) and on mAP (0.5:0.95) with the slightly higher parameters. These results demonstrate that EMA is an efficient module for object detection task, and further proves the effectiveness of the EMA method in this paper.

我们使用YOLOv5x作为我们在VisDrone数据集上进行目标检测的主干卷积神经网络，其中CA、CBAM和EMA注意力分别集成到检测器中。从表2的结果可以看出，CA、CBAM和EMA都能够提升目标检测的基线性能。我们可以看到，我们提出的EMA模块在mAP（0.5）和mAP（0.5:0.95）方面始终优于基于CA和CBAM的基础网络。值得注意的是，CBAM通过 提升了YOLOv5x的性能，并且比 的性能更好，但代价是需要更多的参数和计算。对于CA，它的性能几乎与基线相同，并且在mAP（0.5）方面超越了YOLOv5x ，同时CA的参数和计算量比EMA更高（91.28M 对 91.18M 和 对 ）。具体来说，EMA比基线方法增加了 的参数，并且在mAP（0.5）上比YOLOv5x提高了 ，在mAP（0.5:0.95）上提高了 ，参数量略有增加。这些结果表明EMA是目标检测任务的一个高效模块，并且进一步证明了本文EMA方法的有效性。

Table 4: Ablation on relative training configuration settings on the CIFAR100.

表4：在CIFAR100上的相对训练配置设置的消融研究。

| Method | Datasets | #.Param. | FLOPs | Top-1 (%) | Top-5 (%) |
| --- | --- | --- | --- | --- | --- |
| +EMA\_no | CIFAR100 |  | 1.32G | 78.24 | 94.89 |
| + EMA\_16 |  |  | 80.35 | 95.44 |
| + EMA\_32 |  | 1.32G | 80.69 | 95.59 |

# 5. Ablation Study

# 5. 消融研究

We choose ResNet50 as the baseline network and validate the importance of cross-spatial learning method by conducting ablation experiments to observe the impact of different hyperparameters in EMA, such as EMA\_no (i.e., without no cross-spatial learning), EMA\_16 (i.e., group size is set as 16) and EMA\_32 (i.e., group size is set as 32). Comparing with EMA\_32, a relatively high FLOPs and network parameters will be resulted by setting group size as 16. This is mainly due to reshape the channel dimensions into the batch dimensions that decreases the model parameters. EMA is able to call upon to distribute the model over multiple channels on more batch dimensions and process them. Moreover, we also conduct the ablation study by conducting cross-spatial learning method and the other turns off. From the view of Table 4, EMA\_32 with the cross-spatial learning method outperforms EMA\_no scheme. For the similarly FLOPs and network parameters, the Top-1 and Top-5 rates of EMA\_32 are much higher, at and , respectively.

我们选择ResNet50作为基线网络，并通过进行消融实验来验证跨空间学习方法的重要性，观察EMA中不同超参数（如EMA\_no（即没有跨空间学习），EMA\_16（即组大小设置为16）和EMA\_32（即组大小设置为32））的影响。与EMA\_32相比，将组大小设置为16会导致相对较高的FLOPs和网络参数。这主要是由于将通道维度重塑为批次维度，从而减少了模型参数。EMA能够调用分布在多个通道上的更多批次维度并处理它们。此外，我们还通过启用跨空间学习方法和禁用其他方法进行了消融研究。从表4的角度来看，带有跨空间学习方法的EMA\_32方案优于EMA\_no方案。在相似的FLOPs和网络参数下，EMA\_32的Top-1和Top-5准确率分别高得多，分别为 和 。

# 6. Conclusion

# 6. 结论

In this paper, we systematically investigate the properties of attention mechanisms, which leads to a principled way to combine them into CNNs. Moreover, we present new insight into how the CNNs can enjoy both good generalization and computation budgets by using a generic method that avoids some form of dimensionality reduction via a universal convolution. Due to the flexible and light-weighted characteristics, our proposed EMA can be easily exploited into different computer vision tasks for achieving best performance. We believe our EMA is more applicable to broader applications like semantic segmentation and can be stacked into other deep CNNs structure for significantly enhancing the feature representation ability. We will leave them for future work.

在本文中，我们系统地研究了注意力机制的特性，这导致了一种将它们原理性地结合到CNN中的方法。此外，我们提出了新的见解，即CNNs如何通过使用一种避免某种形式维度降低的通用方法来享受良好的泛化能力和计算预算，该方法通过通用卷积实现。由于我们的EMA具有灵活性和轻量级特性，因此可以轻松地将其应用于不同的计算机视觉任务以实现最佳性能。我们相信我们的EMA更适合更广泛的应用，如语义分割，并且可以堆叠到其他深度CNN结构中，显著增强特征表示能力。我们将把这些留待未来的工作。

# References

# 参考文献

[1] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017.

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, 2018.

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, Jun 2018.

[4] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In ECCV, 2018.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[6] Xiang Li, Xiaolin Hu, and Jian Yang. Spatial groupwise enhance: Improving semantic feature learning in convolutional networks. CoRR, vol. abs/1905.09646, 2019.

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

[8] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate Attention for Efficient Mobile Network Design. In CVPR, 2021.

[9] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized Self-Attention: Towards High-quality Pixel-wise Regression. CoRR, vol. abs/2107.00782, 2021.

[10] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In CVPR, 2020.

[11] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492- 1500.

[12] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang,"Selective kernel networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510-519.

[13] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In CVPR, 2021.

[14] Qing-Long Zhang, and Yu-Bin Yang. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. CoRR, vol. abs/2102.00240, 2021.

[15] Ankit Goyal, Jia Deng, and Vladlen Koltun. Non-deep Networks. CoRR, vol. abs/2110.07641, 2021.

[16] Yichao Liu, Zongru Shao, Yueyang Teng, and Nico Hoffmann. NAM: Normalization-based Attention Module. CoRR, vol. abs/2111.12419, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proc. of Computer Vision and Pattern Recognition (CVPR), 2016.

[18] G. Jocher et al., Ultralytics/YOLOv5: V6.0-YOLOv5n ’nano’ models roboflow integration TensorFlow export OpenCV DNN support, Oct. 2021.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In NeurIPs, 2012.

[20] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In CVPR, 2017.

[21] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. arXiv preprint arXiv:1904.01169, 2019.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016.

[23] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In CVPR, 2019.

[24] Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. EPSANet: An Effificient Pyramid Split Attention Block on Convolutional Neural Network. arXiv:2105.14447[cs.CV], 2021.

[25] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3146- 3154.

[26] Luchen Liu, Sheng Guo, Weilin Huang, and Matthew R Scott, "Decoupling category-wise independence and relevance with self-attention for multi-label image classification," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 1682-1686.

[27] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, "A2̂-nets: Double attention networks," Advances in neural information processing systems, vol. 31, 2018.

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794-7803.

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted

residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510-4520, 2018.

[30] Xingkui Zhu, Shuchang Lyu, Xu Wang and Qi Zhao. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. CoRR, vol. abs/2108.11539, 2021.