Fully Attentional Network for Semantic Segmentation

QiSong, 1, 2 Jie Li, 1, 2*Chenghong Li, *Hao Guo, 1, 2 Rui Huang[1]

[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society
{qisong, jielil, chenghongli, haoguo}@link. cuhk. edu. cn, ruihuang@cuhk. edu. cn

Abstract

Recent non-local self-attention methods have proven to be effective in capturing long-range dependencies for semantic segmentation. These methods usually form a similarity map of$R_{C\times C}$ (by compressing spatial dimensions)or $R_{HW\times HW}$ (by compressing channels)to describe the feature relations along either channel or spatial dimensions, where C is the number of channels, H and W are the spatial dimensions of the input feature map. However, such practices tend to condense feature dependencies along the other dimensions, hence causing attention missing, which might lead to inferior results for small/thin categories or inconsistent segmentation inside large objects. To address this problem, we propose a new approach, namely Fully Attentional Network(FLANet), toencode both spatial and channel attentions in a single similarity map while maintaining high computational efficiency. Specifically, for each channel map, our FLANet can harvest feature responses from allother channel maps, and the associated spatial positions as well, through anovel fully attentional module. Our new method has achieved state-of-the-art performance on three challenging semantic segmentation datasets, i. e., 83. 6%, 46. 99%, and 88. 5% on the Cityscapes test set, the ADE20K validation set, and the PASCAL VOC test set, respectively.

Introduction

Recently, semantic segmentation models achieve great progress by capturing long-range dependencies(Zhao et al. 2017;Yang et al. 2018;Yuan, Chen, and Wang 2020;Sun et al. 2019), in which Non-Local(NL)based methods are the mainstream(Zhao et al. 2018;Fu et al. 2019a;Zhang et al. 2019a;Zhu et al. 2019;Ramachandran et al. 2019). To generate dense and well-rounded contextual information NL based models utilize aself-attention mechanism to ex- plore the interdependencies along the channel (Cao et al. 2019;Zhao et al. 2018)or spatial(Huang et al. 2019;Yin et al. 2020;Song, Mei, and Huang 2021)dimensions. We denote these two variants of NL block as“Channel NL”and “Spatial NL”, respectively, and the architectures of these two variants are illustrated in Fig. 1 (a)and (b). Although

*These authors contributedequally
ICorresponding author

摘要

最近的非局部自注意力方法已被证明可以有效地捕获语义分割的长程依赖关系。这些方法通常形成$R_{C\times C}$(通过压缩空间维度)或$R_{HW\times HW}$(通过压缩通道)的相似度图，以描述沿通道或空间维度的特征关系，其中C是通道数，H和W是输入特征图的空间维度。然而，这种做法往往会沿着其他维度压缩特征依赖关系，从而导致注意力缺失，这可能会导致小/瘦类别的结果较差或大对象内部的分割不一致。为了解决这个问题，我们提出了一种新的方法，即全注意力网络（FLANet），在保持高计算效率的同时，在单一的模拟图像图中编码空间注意力和通道注意力。具体来说，对于每个频道图，我们的 FLANet 可以通过一个新颖的全注意力模块从所有其他频道图中收集特征响应，以及关联的空间位置。我们的新方法在三个具有挑战性的语义分割数据集上取得了最先进的性能，分别在Cityscapes测试集、ADE20K验证集和PASCAL VOC测试集上实现了83. 6%、46. 99%和88. 5%。

介绍

最近，语义分割模型通过捕获长程依赖性取得了长效进展（Zhao et al. 2017;Yang et al. 2018;Yuan、Chen 和 Wang 2020;Sun et al. 2019），其中基于非本地 （NL） 的方法是主流（Zhao et al. 2018;Fu et al. 2019a;Zhang et al. 2019a;Zhu et al. 2019;Ramachandran et，al. 2019 ）。为了生成密集和全面的上下文信息，基于 NL 的模型利用自我注意力机制来挖掘通道沿线的相互依赖性（Cao et al. 2019;Zhao et al. 2018） 或空间 （Huang et al. 2019;Yin et al. 2020;Song， Mei， and Huang 2021 ） 维度。我们将NL块的这两种变体分别表示为“Channel NL”和“Spatial NL”，这两种变体的架构如图1（a）和（b）所示。

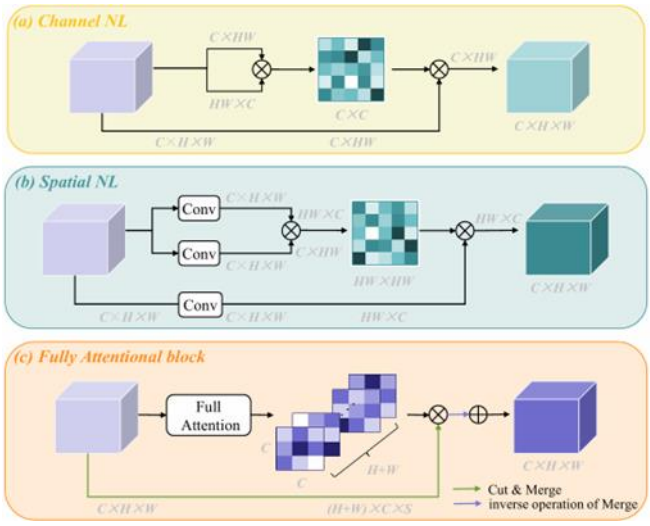


Figure 1:Architectures of Non-Local blocks (NL)and our proposed Fully Attentional block (FLA). The traditional NLs compute a similarity map along channel or spatial dimension, while our FLA generates a similarity map in all dimensions and enables full attentions in the attention map.

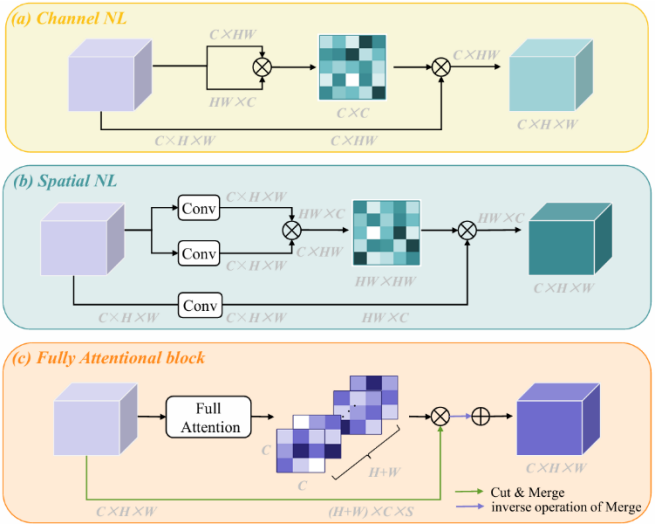


图 1：非本地块 （NL） 的架构和我们提出的全注意力块 （FLA） 。传统的 NL 沿通道或空间差异计算相似性图，而我们的 FLA 生成所有维度的相似性图，并在注意力图中实现全注意力。

these explorationshave made impressive contributions to semantic segmentation, one acute issue, i. e., attention missing, was mostly ignored. Take Channel NL for example, the channel attention map $R_{C\times C}$ is generated by the matrix multiplication of two inputs with a dimension of C×HW and HW×C. It can be found that each channel can be connected with all other channel maps while the spatial information will be integrated and each spatial position fails to perceive feature response from other positions during the matrix multiplication. Similarly, interactions among channel dimensions are also missing in the Spatial NL.

We argue that the attention missing issue would damage the integrity of 3D context information(CHW)and thus both NL variants can only benefit partially in a complementary way. To verify this hypothesis, we present the perclass comparison results on the Cityscapes validation set in Fig. 2. As shown in the figure, Channel NL gets better segmentation results among large objects, such as truck, bus and train, while Spatial NL performs much better on small/thin categories,

虽然这些探索对 SE 语义分割做出了令人印象深刻的贡献，其中一个尖锐的问题，即注意力缺失，大多被忽视了。以通道 NL 为例，通道注意力映射$R_{C\times C}$由两个维度分别为 C×HW 和HW×C的输入矩阵相乘生成. 可以发现，在矩阵乘法过程中，每个通道都可以与所有其他通道映射连接，而空间编队中的空间位置将被整合，并且每个空间位置都无法感知到其他位置的特征响应。同样，在 Spatial NL 中也缺少通道维度之间的交互。我们认为，注意力缺失问题会损害 3D 上下文信息 （CHW） 的完整性，因此两种 NL变体只能以一种完整的方式部分受益。我们认为，注意力缺失问题会损害3D上下文信息（CHW）的完整性，因此两种NL变体只能以互补的方式部分受益。为了验证这一假设，我们在图2中给出了城市景观验证集的每类比较结果。如图所示，Channel NL在卡车、公共汽车和火车等大型物体中获得了更好的分割结，

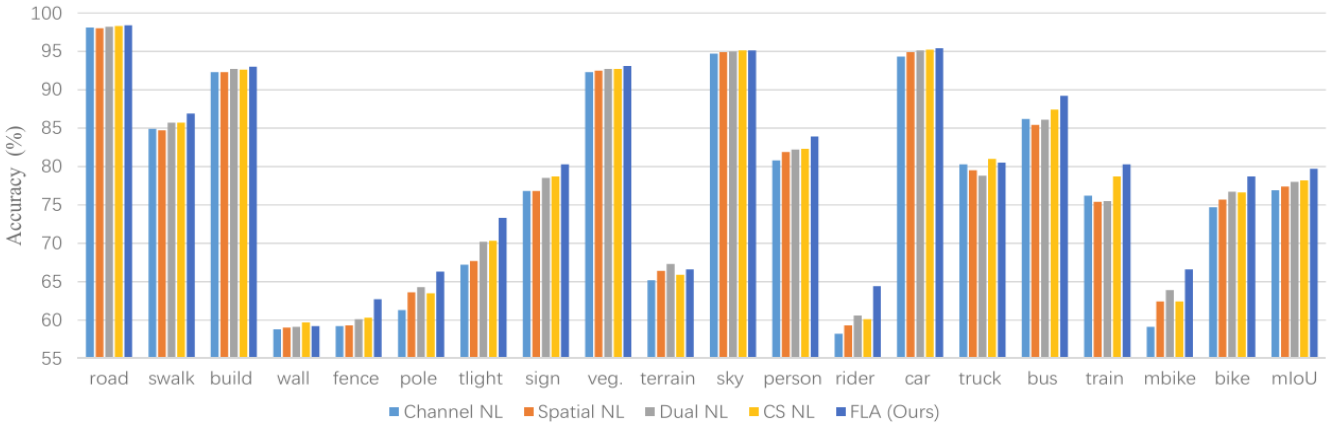


Figure 2:The motivation of our approach and the quantitative evidence for the attention missing issue on the Cityscapes validation set. This observation proves that 1)Spatial NL can enhance the discrimination of details, while Channel NL benefits in maintaining semantic consistency inside largeobjects, 2)stacking two NL blocks in the network still suffers from the attention missing issue, and 3)our proposed FLA can successfully tackle the attention missing issue since we can achieve better accuracy than that of a single NL block in all classes by modeling full attentions.

图 2：我们方法的动机和 Cityscapes 验证集上注意力缺失问题的定量证据。这一观察结果证明：1）空间NL可以增强对细节的判别力，而Channel NL在保持大型对象内部的语义一致性方面具有优势;2）在网络中堆叠两个NL块仍然存在注意力缺失问题;3）我们提出的FLA可以成功地解决注意力缺失问题，因为我们可以通过对全注意力进行建模，在所有类别中达到比单个NL块更好的准确性。

e. g. , poles, rider and mbike. They both lose precision in some categories due to the mentioned attention missing issue. Besides, we are also curious about whether this issue can be solved by stacking the two blocks sequentially We denote the parallel connection mode in DANet (Fu et al. 2019a)and the sequential Channel-Spatial NL as "Dual NL" and "CS NL", respectively[1]. Intuitively, when two NLs are employed at the same time, the accuracy gain of each class should be no less than that of a single NL. However, it is observed that the performance of Dual NL drops a lot in large objects such as truck and train, and CS NL gets poor IoU results in some thin categories like pole and mbike. We can find that both Dual NL and CS NL can only preserve partial benefits brought by either Channel NL or Spatial NL. Therefore, we can conclude that: the attention missing issue hurts the feature representation ability and it cannot be solved by simply stacking different NL blocks.

Motivated by this, we propose a novel non-local block namely Fully Attentional block(FLA)to efficiently retain attentions in all dimensions. And the workflow is shown in Fig. 1(c). The basic idea is to utilize the global context information to receive spatial responses when computing the channel attention map, which enables full attentions in a single attention unit with high computational efficiency. Specifically, we first enable each spatial position to harvest feature responses from the global contexts with the same horizontal and vertical coordinates. Second, we use the self-attention mechanism to capture the fully attentional similarities be- tween any two channel maps and the associated spatial positions. Finally, the generated fully attentional similarities are

'Since there are no convolutional layers in the Channel NL, if we employ Spatial NL before Channel NL(SCNL), the feature weights tend to be either extremely large or extremely small after two consecutive enhancements and the training loss does not converge. So the performance of this connection mode is not reported.

而Spatial NL在电线杆、骑手和共享单车等小/细类别中表现得更好。由于上述注意力缺失问题，它们在某些类别中都失去了精确度。此外，我们也很好奇这个问题是否可以通过按顺序堆叠两个块来解决。SUE可以通过按顺序堆叠两个块来解决。我们将 DANet 中的并联模式（Fu et al. 2019a）和顺序信道空间 NL 分别标记为"双 NL"和"CS NL"[1]。直观地说，当同时使用两个 NL 时，每个类别的精度增益应不小于单个 NL 的精度增益。然而，据观察，Dual NL 在卡车和火车等大型物体上的性能下降了很多，而CS NL在一些薄的类别（如杆子和共享单车）中得到了很差的 IoU 结果。我们可以得出结论，Dual NL 和 CS NL 都只能保留 Channel NL 或 Spatial NL 带来的部分优势。综上所述，我们可以得出结论：注意力缺失问题损害了特征表示能力，并且无法通过简单地堆叠不同的NL块来解决。

受此启发，我们提出了一种新的非局部块，即完全注意块（FLA），以有效地保留所有维度的注意力。工作流程如图1（c）所示。其基本思想是在计算信道注意力图时，利用编队的全局上下文接收空间响应，从而在计算效率高的单注意力单元中实现充分的注意力。具体来说，我们首先使每个空间位置能够从具有相同水平和垂直坐标的全局上下文中获取特征响应。其次，我们使用自注意力机制来捕捉任意两个频道地图之间的完全注意相似性以及相关的空间位置。最后，通过整合所有通道图的特征和相关的全局线索，

由于通道 NL 中没有卷积层，如果我们在通道 NL （SC NL） 之前使用空间 NL，则在连续两次增强后，特征权重往往非常大或非常小，并且训练损失不会收敛。因此，不会报告此连接模式的性能。

used to re-weight each channel map by integrating features among all channel maps and associated global clues.

It should be noted that our method is more effective and efficient than previous works (Fu et al. 2019a;Babiloni et al. 2020)when modeling interdependencies in all dimensions. Since we encode spatial interactions into the traditional Channel NL and capture full attentions in a single attention map, our FLA is with high computational efficiency. Specifically, our FLA significantly reduces FLOPs by about 83%and only requires 34%GPU memory usage of DANet in computing both spatial and channel dependencies.

We have carried out extensive experiments on three challenging semantic segmentation datasets and our approach achieves state-of-the-art performance on these experiments Moreover, our model outperforms other non-local based methods by a large margin with the same backbone network. Our contributions mainly lie in three aspects:

·Through the theoretical and experimental analysis, we find out the attention missing issue existing in the non-local self-attention methods, which would hurt the integrity of feature representation.

·We reformulate the self-attention mechanism into a fully attentional manner to generate dense and well- rounded feature dependencies, which addresses the attention missing issue effectively and efficiently. To the best of our knowledge, this paper is the first to achieve full attentions in a single non-local block.

·We conducted extensive experiments on three chal- lenging semantic segmentation datasets, including Cityscapes, ADE20K, and PASCAL VOC, which demonstrate the superiority of our approach over other state-of-the-art methods.

利用生成的全注意力相似性对每个通道图进行重新加权。

应该指出的是，我们的方法比以前的工作更有效和高效（Fu et al. 2019a;Babiloni et al. 2020），当对所有维度的相互依赖性进行建模时。由于我们将空间交互编码到传统的 Channel NL 中，并在单个注意力图中捕获全部注意力，因此我们的 FLA 具有很高的计算效率。具体来说，我们的 FLA 显著减少了约 83% 的 FLOPs，并且在计算空间和通道依赖性时仅需要 DANet 34% 的 GPU 内存使用率。我们已经在三个具有挑战性的语义分割数据集上进行了广泛的实验，我们的方法在这些实验中取得了最先进的性能。此外，在相同的骨干网络下，我们的模型比其他非本地方法的性能要好得多。我们的贡献主要体现在三个方面：

• 通过理论和实验分析，揭示了非局部自注意力方法中存在的注意力缺失问题，该问题会损害特征表示的完整性。

• 我们将自我注意力机制重新表述为完全注意力的方式，以产生密集且全面的特征依赖关系，从而有效且高效地解决注意力缺失问题。据我们所知，这篇论文是第一个在单个非局部块中实现完全关注的论文。

• 我们对三个具有挑战性的语义分割数据集进行了广泛的实验，包括 Cityscapes、ADE20K 和 PASCAL VOC，这证明了我们的方法优于其他最先进的方法。

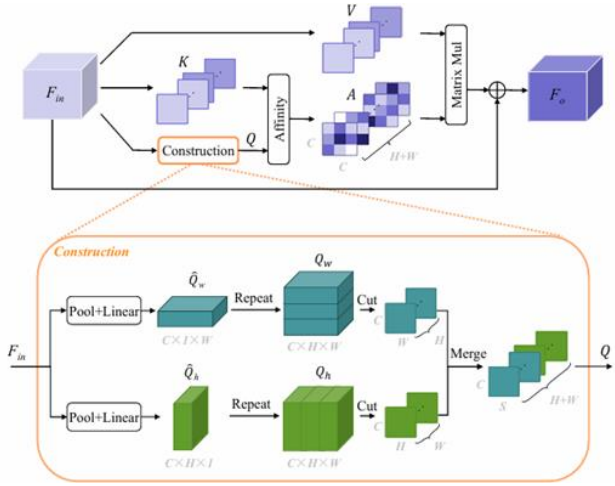


Figure 3:The details of Fully Attentional block. Since H equals W in our implementation, we use the letter S to represent the dimension after merge for a clear illustration.

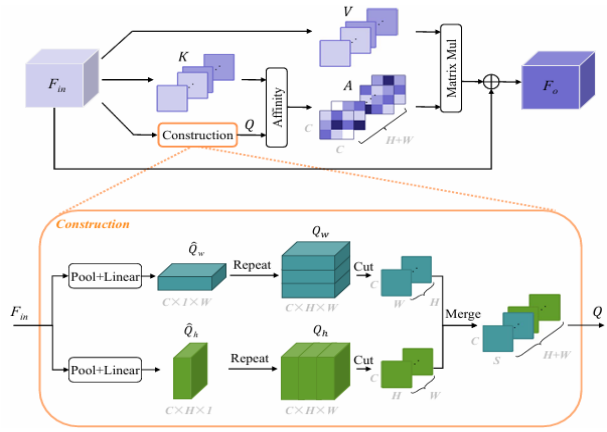


图3：全注意力区块的详细信息。由于在我们的实现中H等于W，为了清楚地说明，我们使用字母S来表示合并后的维度。

Related Work

Semantic Segmentation. Semantic segmentation is a vital task in computer vision, which predicts correct semantic labels for all pixels in an image. The traditional classification network based on CNNs can only identify the class of the whole image, not the label of each pixel. Instead of the fully connected layer in the CNN, the FCN (Long, Shelhamer, and Darrell 2015)utilizes a convolutional layer to get the segmentation result. UNet (Ronneberger, Fischer, and Brox 2015)adopts an encoder-decoder structure to recover the detailed information damaged by the step-by-step downsampling operations. To model interdependencies between different channel maps, SENet (Hu, Shen, and Sun 2018)produces an embedding of the global distribution of channel-wise feature responses. To enhance the global connections between spatial positions, self-attention based methods are thus proposed to weigh the importance of each spatial position whilst sacrificing the channel-wise attention. Different from these approaches, we argue that the attention missing issue might lead to inconsistent segmentation inside large objects or inferior results for small categories in the semantic segmentation task. Thus in this paper, we consider both channel and spatial dependencies are of equal importance and try to capture both of them in a single attention unit.

Self-Attention Mechanism. Self-attention is initially used for machine translation(Chorowski et al. 2015; Vaswani et al. 2017)to capture long-range features. After that, self-attention modules are widely applied in the semantic segmentation field, in which the Non-local network (Wang et al. 2018)is the pioneering work. CCNet (Huang et al. 2019) harvests the contextual information for each pixelon the crisscross path. AttaNet(Song, Mei, and Huang 2021)utilizes a striping operation to encode the global context in the vertical direction and then harvests long-range relations along the horizontal axis. OCNet (Yuan et al. 2021)

相关工作

语义分割。语义分割是计算机视觉中的一项重要任务，它可以预测图像中所有像素的正确语义标签。传统的基于 CNN 的分类网络只能识别整个图像的类别，而不能识别每个像素的标签。FCN（Long， Shelhamer， and Darrell 2015）利用卷积层来获得分割结果，而不是 CNN 中的全连接层。UNet（Ronneberger， Fischer，and Brox 2015）采用编码器-解码器结构来恢复因逐步下采样操作而损坏的详细信息。为了对不同通道图之间的相互依赖性进行建模，SENet（Hu, Shen, and Sun 2018）生成了通道特征响应的全局分布的嵌入。为了增强空间位置之间的全局联系，该文提出基于自我注意力的方法，在牺牲通道注意力的同时权衡每个空间位置的重要性。与这些方法不同，我们认为注意力缺失问题可能会导致大对象内部的分割不一致，或者在语义分割任务中导致小类别的结果较差。因此，在本文中，我们认为通道和空间依赖性具有同等的重要性，并尝试在一个注意力单元中捕获它们。

自我关注机制。 自我关注最初用于机器翻译（Chorowski et al. 2015;Vaswani et al. , 2017）来捕获远距离特征。此后，自注意力模块在语义分割领域得到广泛应用，其中非局域网络（Wang et al. 2018）是其中的开创性工作。CCNet （Huang et al. 2019）在纵横交错的路径上收集每个像素的上下文信息。AttaNet（Song，Mei，and Huang 2021）利用条带化运算在垂直方向上对全局文本进行编码，然后沿水平轴收集长距离关系。

utilizes the interlaced self-attention scheme to model both global and local relations. However, these approaches construct a similarity map to leverage relationships along a single dimension, where the dependency along other dimensions is discarded during the matrix multiplication. To generate both spatial-wise and channel-wise attentions, many studies were proposed. DANet(Fu et al. 2019a) proposes the position attention module and channel attention module to model dependencies along the spatial and channel dimension respectively. TESA(Babiloni et al. 2020)views the input tensor as a combination of its three-mode matricizations and then captures similarities for each dimension. Although these methods capture relations in all dimensions, they consider different dimensions separately and the attention missing issue still exists in each attention map. To mitigate this issue, we propose a Fully Attentional block to encode both spatial and channel attentions in a single similarity map with high computational efficiency.

Method

Network Architecture

In this paper, we employ ResNet101(He et al. 2016)and HRNetW48(Sun et al. 2019)as the backbone network. For ResNet101, dilation convolutions were applied in the last two layers to obtain more detailed information, and the output feature map was enlarged to 1/8 of the input image. Initially, the input image is processed by the backbone network to produce feature maps X. After that, we first apply two convolution layers on X to reduce the channel dimension and obtain the feature maps F_{in} Then, the feature maps F_{in} would be fed into the Fully Attentional block (FLA)and generate new feature maps F_o which aggregate non-local contextual information in all dimensions. Finally, the dense contextual feature F_o is sent to the prediction layer to generate the final segmentation map.

Fully Attentional Block

Previous works try to generate full attentions by applying the attention operation in each dimension in turn, which yields high computational complexity and the single-dimensional attention still overlooks correlations along other dimensions. To capture full attentions in a single attention map with high computational eficiency, we propose anovel non-local block named Fully Attentional block. Specifically, to avoid adding extra computation burden, we try to introduce spatial interactions into channel NL mechanism by utilizing the global average pooling result as the global contextual prior.

The pipeline of our method is shown in Fig. 3. Given an input feature map $F_{in} \in R^{C\times H\times W,}$ where C is the number of channels, H and W are the spatial dimensions of the input tensor. First, we feed F_{in} into two parallel pathways at the bottom (i.e., the Construction**)**, each of which contains a global average pooling layer followed by a Linear layer. When choosing the size of pooling windows, we considered the following two aspects. Firstly, to obtain richer global contextual priors, we choose to use unequal global pooling size in height and width directions rather than kernel windows like 3×3.

利用交错自注意力方案对全球和局部关系进行建模。然而，这些方法构建了一个相似性图，以利用沿单一维度的关系，其中沿其他维度的依赖性在矩阵乘法过程中被丢弃。为了在空间和通道上引起关注，提出了许多研究。DANet（Fu et al. 2019a）提出了位置注意力模块和通道注意力模块，分别对空间维度和通道维度的依赖关系进行建模。TESA（Babiloni et al. 2020）将输入张量视为其三模矩阵化的组合，然后捕获每个维度的相似性。尽管这些方法捕获了所有维度的关系，但它们分别考虑了不同的维度，并且每个注意力图中仍然存在注意力缺失问题。为了缓解这个问题，我们提出了一个完全注意块，以提高计算效率在单个相似性图中编码空间和通道注意力。

方法

网络架构

本文采用 ResNet101（He et al. 2016）和 HRNetW48（Sun et al. 2019）作为骨干网络。对于 ResNet101，在最后两层应用膨胀卷积以获得更详细的信息，并将输出的特征图放大到输入图像的 1/8。首先，输入图像由骨干网络处理生成特征图 X. 然后，首先在 X 上应用两个卷积层来减小通道维数并得到特征图F_{in}. 然后，将特征图F_{in}输入到完全注意力块（FLA）中并生成新的特征图F_o它聚合了所有维度的非本地上下文信息。最后，对密集上下文特征F_o表示。被发送到预测层以生成最终的分割图。

完全注意力阻滞

以往的研究试图通过在各个维度依次应用注意力操作来产生充分的注意力，这产生了较高的计算复杂度，而一维的注意力仍然忽略了其他维度的相关性。为了在单个注意力图中具有较高的计算效率捕获全部注意力，我们提出了一种名为完全注意块的新型非局部块。具体而言，为了避免增加额外的计算负担，我们尝试利用全球平均池化结果作为全局上下文先验，在通道 NL 机制中引入空间交互。

我们方法的流水线如图3所示。给定输入特征图$F_{in} \in R^{C\times H\times W}$，其中C是通道的数量，H和W是输入张量的空间维度。首先，我们将F_{in}输入底部的两个平行路径（即构造），每个路径都包含一个全局平均池化层和一个线性层。在选择池窗口的大小时，我们考虑了以下两个方面。首先，为了获得更丰富的全局上下文先验，我们选择在高度和宽度方向上使用不相等的全局池大小，而不是像3×3这样的内核窗口。

Secondly, to make sure that each spatial position is connected with the corresponding global prior with the same horizontal or vertical coordinate, i. e., maintain the spatial consistency when computing channel relations, we choose to keep the length of one dimension constant. Therefore, we employ pooling windows of size H×1 and 1×W in these two pathways respectively. This gives $Q_w \in R^{C\times 1\times W}$ and $Q_h \in R^{C\times H\times 1}$. After that, we repeat Q_w and Q_h to form global features $Q_w \in R^{C\times H\times W}$ and $Q_h \in R^{C\times H\times W}$. Note that Q_w and Q_h represent the global priors in the horizontal and vertical directions respectively and they will be used to achieve spatial interactions in the corresponding dimension. Furthermore, we cut Q_w along the H dimension, from which we can generate a group of H slices with a size of $R^{C\times W}$. Similarly, we cut Q_h along the W dimension. We then merge these two groups to form the final global contexts $Q \in R^{(H+W)\times C\times S}$. The cut and merge operations are detailly illustrated in Fig. 3.

Meanwhile, we cut the input feature F_{in} along the H dimension, yielding a group of H slices with the size of $R^{C\times W}$. Similarly, we do this along the W dimension. Like the merge process of Q, these two groups are integrated to form the features $K \in R^{(H+W)\times S\times C}$. In the same way, we can generate the feature maps $V \in R^{(H+W)\times C\times S}$.

After that, we can make each spatial position to receive the feature responses from the global priors in the same row and the same column, i. e., capturing the full attentions $A \in R^{(H+W)\times C\times C}$, via the Affinity operation. The Affinity operation is defined as follows:

$$A_{i,j} = \frac{exp(Q_i \cdot K_j)}{\sum_{i=1}^{C} exp(Q_i \cdot K_j)} \tag{1}$$

where $A_{i,j} \in A$ denotes the degree of correlation between the i_{th} and j_{th} channel at a specific spatial position.

Then we perform amatrix multiplication between A and V to update each channel map with the generated full attentions. After that, we reshape the result into two groups and each group is with a size of $R^{C\times H\times W}$(i. e., the inverse operation of merge). We sum these two groups to form the long-range contextual information. Finally, we multiply the contextual information by a scale parameter γ and perform an element-wise sum operation with the input feature map F_{in} to obtain the final output $F_o \in R^{C\times H\times W}$ as follows:

$$F_{o_j} = \gamma \sum_{i=1}^{C} A_{i,j} \cdot V_j + F_{in_j}, \tag{2}$$

where F_o is a feature vector in the output feature map F, at the j_{th} channel map.

It is noted that different from the traditional Channel NL method which explores only channel correlations by multiplying the spatial information from the same position, our FLA enables spatial connections between different spatial positions, i.e., we exploit full attentions along both spatial and channel dimensions with a single attention map. In this way, our FLA has a more holistic contextual view and is more robust to different scenarios. Moreover, the constructed prior representation brings a global receptive field and helps to boost the feature discrimination ability.

其次，为了确保每个空间位置与具有相同水平或垂直坐标的相应全局先验相连接，即在计算通道关系时保持空间一致性，我们选择保持一维长度不变。因此，我们在这两个路径中分别使用大小为H×1和1×W的池化窗口。这给出了$Q_w \in R^{C\times 1\times W}$和 $Q_h \in R^{C\times H\times 1}$。然后，我们重复 Q_w 和 Q_h，形成全局特征 $Q_w \in R^{C\times H\times W}$和 $Q_h \in R^{C\times H\times W}$。注意，$Q_w$和$Q_h$分别表示水平和垂直方向上的全局先验，它们将用于实现相应维度上的空间交互。此外，我们沿着H维度切割Q_w，从中我们可以生成一组大小为$R^{C\times W}$的H切片。类似地，我们沿着W维度切割Q_h。然后，我们将这两个群合并，形成最终的全局上下文$Q \in R^{(H+W)\times C\times S}$。剪切和合并操作如图3所示。

同时，我们沿着H维度切割输入特征F_{in}，得到一组大小为$R^{C\times W}$的H切片。类似地，我们沿着W维度这样做。像Q的合并过程一样，这两个群被积分形成特征$K \in R^{(H+W)\times S\times C}$。同样，我们可以生成图3中的特征映射$V \in R^{(H+W)\times C\times S}$。

然后，我们可以使每个空间位置接收来自同一行同一列的全局先验的特征响应，即通过仿射运算捕获全部注意力 $A \in R^{(H+W)\times C\times C}$。仿射运算定义如下：

$$A_{i,j} = \frac{e^{Q_i \cdot K_j}}{\sum_{i=1}^{C} e^{Q_i \cdot K_j}}$$

其中$A_{i,j} \in A$表示在特定的空间位置上的第i和第j通道。

然后我们在A和V，以使用生成的完整更新每个通道图保持。之后，我们将结果重新定义为两组，并且每个组的大小为$R^{C\times H\times W}$（即逆合并操作）。我们将这两组相加，形成长期上下文信息。最后，我们将上下文信息通过比例参数γ 和执行对输入特征图的逐元素求和运算F_{in}得到的最终输出$F_o \in R^{C\times H\times W}$如下：

$$F_{oj} = \gamma \sum_{i=1}^{C} A_{i,j} \cdot V_j + F_{in\,j}$$

其中F_o，是输出特征图F中的特征向量，在第 j 个频道地图上。

结果表明，与传统的通道 NL方法仅通过相乘来自同一位置的空间信息来探索通道相关性不同，我们的FLA实现了不同空间位置之间的空间联系，即通过单一的注意力图在空间维度和通道维度上都充分挖掘注意力。这样一来，我们的 FLA 就有了更全面的上下文视图，对不同场景的鲁棒性更强。此外，构造的先验表示带来了全局感受场，有助于提高特征辨别能力。

Method	Backbone	mloU
Simple Backbone		
PSPNet(Zhao et al. 2017)	Res101	78. 4
AAF(Ke et al. 2018	Res101	79. 1
CFNet (Zhang, Wang, and Xie 2019)	Res101	79. 6
PSANet (Zhao et al. 2018)	Res101	80. 1
AttaNet (Song, Mei, and Huang 2021)	Res10	80. 5
ANNet (Zhu et al. 2019)	Resl0	81. 3
CCNet (Huang et al. 2019)	Res10	81. 4
OCNet(Yuan et al. 2021)	Res10	81. 9
DGCNet(Zhang et al. 2019b)	Res101	82. 0
HANet(Choi, Kim, and Choo 2020)	Res101	82. 1
ACNet(Fu et al. 2019b)	Res101	82. 3
RecoNet(Chen et al. 2020)	Res10	82. 3
FLANet (Ours	Res101	83. 0
Advanced Backbone		
SPGNet(Cheng et al. 2019)	2×Res50	81. 1
DANet(Fu et al. 2019a)	R101+MG	81. 5
ACFNet(Zhang et al. 2019a)	R101+ASPP	81. 8
GALD(Liet al. 2019)	R101+ASPP	81. 8
GFF (Liet al. 2020)	R101+PPM	82. 3
HRNet(Sun et al. 2019)	HRNetW48	81. 6
OCNet (Yuan et al. 2021)	HRNetW48	82. 5
FLANet (Ours)	HRNetW48	83. 6

Table 1:Comparison with state-of-the-art models on the Cityscapes test set. For fair comparison, all these methods use only Cityscapes fine-data for training.

方法	骨干网	mloU
简易骨干		
PSPNet(Zhao et al. 2017)	Res101	78. 4
AAF(Ke et al. 2018	Res101	79. 1
CFNet (Zhang, Wang, and Xie 2019)	Res101	79. 6
PSANet (Zhao et al. 2018)	Res101	80. 1
AttaNet (Song, Mei, and Huang 2021)	Res10	80. 5
ANNet (Zhu et al. 2019)	Resl0	81. 3
CCNet (Huang et al. 2019)	Res10	81. 4
OCNet(Yuan et al. 2021)	Res10	81. 9
DGCNet(Zhang et al. 2019b)	Res101	82. 0
HANet(Choi, Kim, and Choo 2020)	Res101	82. 1
ACNet(Fu et al. 2019b)	Res101	82. 3
RecoNet(Chen et al. 2020)	Res10	82. 3
FLANet (Ours	Res101	83. 0
先进骨干		
SPGNet(Cheng et al. 2019)	2×Res50	81. 1
DANet(Fu et al. 2019a)	R101+MG	81. 5
ACFNet(Zhang et al. 2019a)	R101+ASPP	81. 8
GALD(Liet al. 2019)	R101+ASPP	81. 8
GFF (Liet al. 2020)	R101+PPM	82. 3
HRNet(Sun et al. 2019)	HRNetW48	81. 6
OCNet (Yuan et al. 2021)	HRNetW48	82. 5
FLANet (Ours)	HRNetW48	83. 6

表1：与Cityscapes测试集上的最新模型的比较。为了公平比较，所有这些方法都只使用Cityscapes的精细数据进行训练

Complexity Analysis

Given a feature map with a size of C×H×W, the typical Spatial NL has a computational complexity of O((HW)²C), and the Channel NL has a computational complexity of O(C²HW). Both of them can only capture similarities along a single dimension. To model feature dependencies in all dimensions, previous work like DANet applies both Spatial NL and Channel NL to calculate spatial and channel relations separately, which yields higher computational complexity and occupies much more GPU memory. Different from previous works, we achieve full attention in a single NL block and in a more efficient way. Specifically, we utilize the newly constructed global representations to achieve interactions between different spatial positions and collect contextual similarities from all dimensions. And the complexity of our FLA block (both in time and space) is O(C²(H+W)S). Since S=H=W in our paper, our complexity is of the same order with Channel NL and only differs by a small constant.

Experiments

To evaluate the proposed FLANet, we conduct extensive experiments on the Cityscapes (Cordts et al. 2016), the ADE20K(Zhou et al. 2017), and the PASCAL VOC (Everingham et al. 2009).

Datasets

Cityscapes Cityscapes is a dataset for urban scene segmentation, which contains 5K images with fine pixel-level

复杂性分析

给定大小为C×H×W的特征图，典型的空间NL的计算复杂度为O（（HW）²C），通道NL的计算复杂度为O（C²HW）。它们都只能沿单一维度捕获相似性。为了对所有维度的特征变化进行建模，前人的工作如DANet同时应用了Spatial NL和Channel NL来分别计算空间关系和通道关系，这产生了更高的计算复杂度，占用了更多的GPU内存。与以往的工作不同，我们以更有效的方式在单个NL块中实现了全神贯注。具体而言，利用新构建的全局表示实现不同空间位置之间的交互，并从各个维度收集上下文相似性。我们的FLA块的复杂度（无论是在时间和空间上）都是O（C²（H+W）S）。由于本文中的S=H=W，我们的复杂度与信道NL相同，只是相差一个小常数。

实验

为了评估所提出的FLANet，我们对城市景观（Cordts et al. 2016）、ADE20K（Zhou et al. 2017）和PASCAL VOC（Everingham et al. 2009）进行了广泛的实验。

Cityscapes Cityscapes 是一个用于城市场景分割的数据集，其中包含具有精细像素级的5K图像

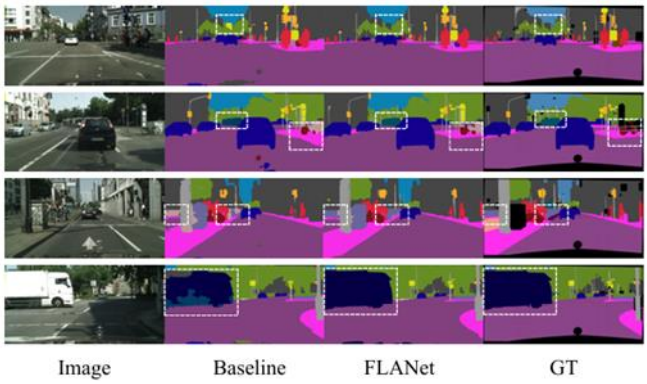


Figure 4:Visualization results of FLANet on the Cityscapes validation set.

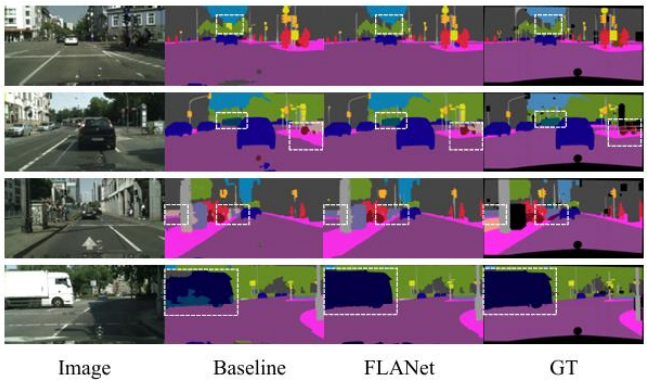


图 4：FLANet 在 Cityscapes 验证集上的可视化结果。

annotations and 20K images with coarse annotations. The dataset has 19 classes and each image is with 1024×2048 resolution. The 5K images with fine annotations are further divided into 2975, 500, and 1525 images for training, validation, and testing, respectively.

ADE20K ADE20K is a challenging scene parsing benchmark. The dataset contains 20K/2K images for training and validation which is densely labeled as 150 stuff/object categories. Images in this dataset are from different scenes with more scale variations.

PASCAL VOC PASCAL VOC is a golden benchmark of semantic segmentation, which includes 20 object categories and one background class. The dataset contains 10582, 1449, 1456 images for training, validation, and testing.

Implementation Details

Our implementation is based on PyTorch, and uses ResNet101 and HRNetW48 pre-trained from ImageNet (Russakovsky et al. 2015) as the backbone network. Following prior works(Yu et al. 2018), we apply the poly learning rate policy where the initial learning rate is multiplied by $\left(1-\frac{iter}{max_{iter}}\right)^{0.9}$after each iteration. Momentum and weight decay coefficients are set to 0. 9 and 5e-4, respectively. All models are trained for 240 epochs with an initial learning rate of 1e-2 and batch size of 8. We set the crop size as 768×768 and 520×520 for Cityscapes and other data sets, respectively. For data augmentation, we apply the common color jittering, scaling (0. 5 to 2. 0), cropping, and flipping to augment the training data. Besides, the synchronized batch normalization is used to synchronize the mean and standard deviation of batch normalization across multiple GPUs. For evaluation, the commonly used Mean IoU metric is adopted.

Experiments on the Cityscapes Dataset

Comparisons to the State of the Art We first compare our proposed method with the state-of-the-art approaches on the Cityscapes test set. Specifically, all models are trained with only fine annotated data, and the comparison results

注释和带有粗略注释的 20K 图像。该数据集有 19 个类别，每张图像的分辨率为 1024×2048。将带有精细标注的5K图像进一步分为2975, 500张和1525张图像，分别用于训练、评估和测试。

ADE20K ADE20K是一个具有挑战性的场景解析基准。该数据集包含 20K/2K 图像用于训练和验证，这些图像被密集标记为 150 个事物/对象类别。此数据集中的图像来自不同的场景，具有更多的尺度变化。

PASCAL VOC PASCAL VOC 是语义分割的黄金基准，它包括 20 个对象类别和一个背景类。数据集包含 10582，1449, 1456 张图像，用于训练、验证和测试。

实验细节

ResNet101 和HRNetW48 从ImageNet（Russakovsky et al. 2015）作为骨干网络进行预训练。根据先前的工作（Yu et al. 2018），我们应用多学习率策略，其中每次迭代后初始学习率乘以$\left(1-\frac{iter}{max_{iter}}\right)^{0.9}$。动量衰减系数和重量衰减系数分别设置为0. 9和5e-4。所有模型都被训练了240个时期，初始学习率为1e-2，批量大小为8。我们将城市景观和其他数据集的裁剪尺寸分别设置为768×768和520×520。对于数据增强，我们应用常见的颜色抖动、缩放（0. 5到2. 0）、裁剪和翻转来增强训练数据。此外，同步的批量归一化用于同步多个GPU之间的批量归一化的平均值和标准差。为了进行评估，采用了常用的Mean IoU度量。

在 Cityscapes 数据集上的实验

与最新技术的比较 我们首先将我们提出的方法与Cityscapes测试集上的最新方法进行比较。具体来说，所有模型都只使用精细的注释数据进行训练，并且将比较结果

Method	SS(%)	MS+F(%)
ShuffleNetV2	69. 2	70. 8
+FLA	74. 7	76. 3
Res18	71. 3	72. 5
+FLA	76. 5	78. 1
Res50	72. 8	74. 1
+FLA	78. 9	79. 7
Res101	75. 6	76. 9
+FLA	81. 3	82. 2

Table 2:Ablation study between the baseline and FLANet on Cityscapes validation set according to various backbone networks. SS: Single scale input during evaluation. MS: Multi-scale input. F: Adding left-right flipped input.

方法	SS(%)	MS+F(%)
随机 NetV2	69. 2	70. 8
+FLA	74. 7	76. 3
Res18	71. 3	72. 5
+FLA	76. 5	78. 1
Res50	72. 8	74. 1
+FLA	78. 9	79. 7
Res101	75. 6	76. 9
+FLA	81. 3	82. 2

表2：根据各种骨干网络在Cityscapes验证集上的基线和FLANet之间的消融研究。SS：评估期间的单刻度输入。MS：多尺度输入。F：增加左右翻转输入。

are summarized in Tab. 1. Among these approaches, the self-attention based models are most related to our method, and more detailed analyses and comparisons will be illustrated in the following subsection.

From Tab. 1, it can be observed that our approach substantially outperforms all the previous techniques based on ResNet101(Res101)or stronger backbones and achieves a new state-of-the-art performance of 83. 6% mIoU. Moreover, it achieves a performance that is comparable with methods based on some larger backbones.

Ablation Studies To demonstrate the wide applicability of FLANet, we conduct ablation studies on various back- bone networks, including ShuffleNetV2(Maet al. 2018) and ResNet series. As listed in Tab. 2, models with FLA consistently outperform baseline models with significant increases no matter what backbone network we use.

In addition, we provide the qualitative comparisons between FLANet and the Baseline (ResNet50) in Fig. 4, where we use the white squares to mark the challenging regions One can observe that the baseline easily misclassifies those regions but our proposed network is able to correct them. For example, building in row 1, side walk and distant train in row 2, and large truck in row 4. It also proves the benefits of the proposed FLANet in predicting distant objects and maintaining the segmentation consistency inside large objects.

Comparison with NLMethods We compare our FLANet with several existing non-local models on the Cityscapes validation set. We measure the increased computation complexity (measured by the number of GFLOPs) and GPU memory usage that are introduced by the NL blocks and do not count the complexity from the baselines. Besides, to speed up the training procedure, we carry out these comparison experiments on ResNet50, with batch size 8.

Specifically, the NL models compared in Tab. 3 include 1) Expectation-Maximization Attention in EMANet (Li et al. 2019), donated as "+ EMA" ; 2) Recurrent Criss-Cross Attention(R= 2) in CCNet (Huang et al. 2019), donated as "+RCCA";3)two typical NL blocks introduced in Sec. 1., donated as "+Channel NL" and "+Spatial NL" respectively; 4) two connection modes introduced in Sec. 1., donated as

概括在这些方法中。基于自我注意的模型与我们的方法关系最为密切，更详细的分析和比较将在以下小节中进行说明。

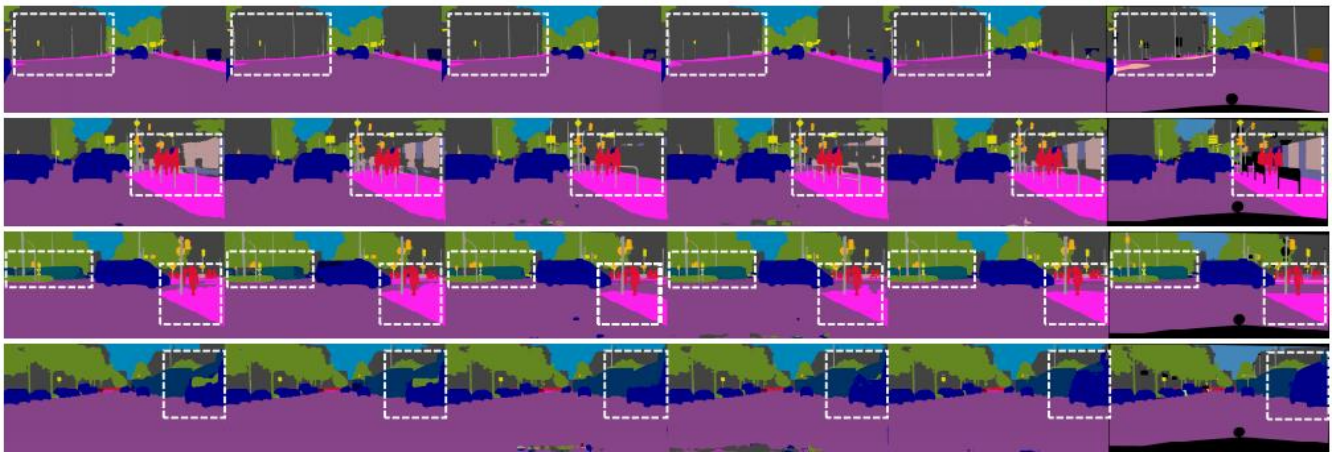
表1可以看出，我们的方法大大优于之前所有基于ResNet101（Res101）或更强骨干网的技术，并取得了83. 6% mIoU的新性能。

消融研究 为了证明 FLANet 的广泛适用性，我们对各种骨干网络进行了消融研究，包括ShuffleNetV2（Maet al. 2018）和 ResNet 系列。如表2所示，无论我们使用何种骨干网络，具有FLA的模型都明显优于基线模型，并且有显著的提高。

此外，我们还提供了 FLANet 和图 4 中的基线(ResNet50) 之间的定性比较，其中我们使用白色方块来标记具有挑战性的区域，人们可以观察到基线很容易对这些区域进行错误分类，但我们提出的网络能够纠正它们。例如，第1排建筑，第2排人行道和远处的火车，第 4. It 排的大货车，也证明了所提出的 FLANet 在预测远处物体和保持大物体内部分割一致性方面的优势。

与 NLMethods 的比较 我们将我们的 FLANet 与 Cityscapes 验证集上的几个现有非本地模型进行了比较。我们测量了 NL 模块引入的增加的计算复杂度（通过 GFLOP 数量衡量）和 GPU 内存使用率，并且不计算基线的复杂度。此外，为了加快训练过程，我们在批量为 8 的 ResNet50 上进行了这些比较实验。

具体而言，表3中比较的NL模型包括：1）EMANet中的期望最大化注意力（Li et al. 2019），以"+ EMA"表示;2 ） CCNet （ Huang et al. 2019） 中的反复交叉注意力 （R= 2 ），捐赠为 "+RCCA";3）Sec. 1. 中介绍的两个典型NL块，分别捐赠为 "+Channel NL"和"+Spatial NL"; 4）第1节中介绍的两种连接方式，捐赠为



Channel NL Spatial NL Dual NL CS NL FLA(Ours) GT

Figure 5:Qualitative comparisons against the NL methods on the Cityscapes validation set. Due to the limited space, we remove the input images and only show the segmentation results and ground truth(GT).

图 5：在 Cityscapes 验证集上与 NL 方法的定性比较。由于空间有限，我们删除了输入图像，只显示分割结果和地面实况（GT）。

Method	SS(%)	MS+F(%)	GFLOPs	Memory
Baseline	72. 8	74. 1		
+EMA	76. 4	77. 1	13. 91	98
+Channel NL	75. 6	76. 9	**9. 66**	**40**
+RCCA	76. 7	77. 8	16. 18	174
+Spatial NL	76. 3	77. 6	103. 90	1320
+Dual NL	77. 1	78. 0	113. 56	1378
+CS NL	77. 4	78. 2	113. 56	1378
+FLA(Ours)	**78. 9**	**79. 7**	19. 37	436

Table 3:Detailed comparisons with existing NL models on the Cityscapes validation set. The GFLOPs and Memory are computed with the input size 768×768. Adding FLA to the baseline largely increase the mIoU with fewer computation. "+Dual NL"and"+CS NL" respectively. Besides, according to whether calculate the channel-only attention, spatial-only attention, and both channel-spatial attention, Tab. 4 is divided into three groups

As illustrated in Tab.3, FLA outperforms these NL methods by a large margin, and the complexity comparison results indicate that the cost of adding FLA is practically negligible even compared with the lightweight-designed models like EMA and RCCA. Moreover, it can be found that the increased computational cost of FLA for capturing spatial attentions is the lowest(about 9. 71 GFLOPs)compared with all these spatial-modeling NLs. Even when compared with the Channel NL who requires the lowest computational cost, our FLA outperforms it by 2. 8%with the minimum computational increment. And the computational complexity of FLA is consistent to our previous theoretical analysis in Sec. 3. 3. It is noted that our FLA significantly reduces GFLOPs by about 83%and only requires 34%GPU memory usage of DANet (Dual NL)and CS NL when modeling

方法	SS(%)	MS+F(%)	GFLOPs	记忆
基线	72. 8	74. 1		
+EMA	76. 4	77. 1	13. 91	98
+Channel NL	75. 6	76. 9	**9. 66**	**40**
+RCCA	76. 7	77. 8	16. 18	174
+Spatial NL	76. 3	77. 6	103. 90	1320
+Dual NL	77. 1	78. 0	113. 56	1378
+CS NL	77. 4	78. 2	113. 56	1378
+FLA(Ours)	**78. 9**	**79. 7**	19. 37	436

表 3：与 Cityscapes 验证集上现有 NL 模型的详细比较。GFLOP 和内存的计算输入大小为 768×768。将 FLA 添加到基线会大大增加 mIoU，但计算量会减少。

分别为"+Dual NL"和"+CS NL"。此外，根据是否计算仅通道注意力、仅空间注意力和两个通道空间注意力，将表4分为3组

如表3所示，FLA的性能大大优于这些NL方法，复杂性比较结果表明，即使与EMA和RCCA等轻量级设计的模块相比，添加FLA的成本实际上也是微不足道的。此外，可以发现，与所有这些空间建模NLs相比，FLA捕获空间注意力的计算成本最低（约为9. 71 GFLOPs），即使与计算成本要求最低的Channel NLs相比，在最小计算增量下，我们的FLA性能也比其高出2. 8%。FLA的计算复杂度与我们之前的理论分析一致，Sec. 3. 3. It 注意到，我们的FLA显著降低了约83%的GFLOPs，并且只需要34%的GPU内存使用DANet（Dual NL）和CS NL在建模

both channel-wise and spatial-wise relationships. Therefore FLA has a great advantage of not only an effective way of improving segmentation accuracy but also a lightweight algorithm design for practical usage.

The Efficacy of FLA To further prove that our method can successfully solve the attention missing issue, we also present several qualitative comparison results in Fig. 5. As shown in Fig. 5, we can find that Dual NL and CS NL can combine the advantages of Channel NL and Spatial NL to some extent and generates better segmentation results. However, it is obvious that sometimes they obtain wrong predictions even when they are correctly classified in Channel NL and Spatial NL, such as the examples shown in the second row. This coincides with our claims that the attention missing issue would distort interactions between dimensions and can not be solved by stacking different NL blocks. Compared with those NL methods, accuracies of predictions for both distant/thin categories (e. g., poles in the first row) and the ability to maintain the segmentation consistency inside large objects (e. g., train in the third row and car in the last row)are significantly improved after using the proposed FLA. And the quantitative per-class comparisons can be seen in Fig. 2. This phenomenon can also demonstrate that FLA can optimally model both channel-wise and spatial-wise relations only by using a single non-local block.

Visualization of Attention Module To get a deeper understanding of the effectiveness of our FLA in encoding spatial attentions into the channel affinity map, we visualize the attended feature maps and analyze how FLA improves the final result. We also visualized the attended feature maps of the traditional Channel NL for further comparison. As shown in Fig. 6, both Channel NL and our FLA highlight some semantic areas and guarantee consistent representation inside large objects like roads and buildings. Furthermore, it is noted that the attended feature maps of FLA are more structured and detailed than that of Channel NL. For example,

通道关系和空间关系时。因此，FLA具有很大的优势，不仅能有效提高分割精度，而且能进行轻量级的算法设计，便于实际应用。

FLA的功效 为了进一步证明我们的方法能够成功地解决注意力缺失问题，我们还给出了几个定性比较 Fig. 5. As 结果，如图5所示，我们发现Dual NL和CS NL可以在一定程度上结合Channel NL和Spatial NL的优点，产生更好的分割结果。但是，很明显，即使它们在通道 NL 和空间 NL 中被正确分类，有时它们也会获得错误的预测，例如第二行中显示的示例。这与我们的说法不谋而合，即注意力缺失问题会扭曲维度之间的交互，并且无法通过堆叠不同的 NL 块来解决。与NL方法相比，使用所提出的FLA方法后，对远/薄类别（例如，第一行的极点）的预测精度和保持大型目标内部分割一致性的能力（例如，火车在第三排，汽车在最后一排）均有显著提高。这种现象也可以证明，FLA仅通过使用单个非局部块就可以对通道和空间关系进行最优建模。

注意力模块的可视化 为了更深入地了解我们的 FLA 在将空间注意力编码到通道亲和力图中的有效性，我们将有人值守的特征图可视化，并分析 FLA 如何改善最终结果。我们还可视化了传统 Channel NL 的有人值守特征图，以便进一步比较。如图 6 所示，Channel NL 和我们的 FLA 都突出了一些语义区域，并保证了在道路和建筑物等大型物体内的一致表示。此外，需要注意的是，FLA的有人值守特征图比Channel NL的特征图更有条理和详细。例如，

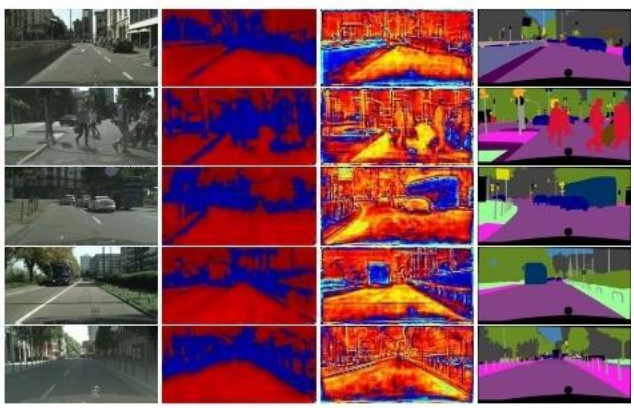


Figure 6:Visualization of attended feature maps of Channel NL and our FLA on the Cityscapes validation set, where feature maps are visualized by averaging along the channel dimension.

distant poles and object boundaries are highlighted for all images. Particularly, FLA can also distinguish different classes, e. g., bus and car in the third row. These visualization results further demonstrate that our proposed module can capture and encode the spatial similarities into the channel attention map to achieve full attentions.

Experiments on the ADE20K Dataset

To further validate the effectiveness of our FLANet, we conduct experiments on the ADE20K dataset, which is a challenging scene parsing dataset with both indoor and outdoor images. Tab. 4 reports the performance comparisons between FLANet and the state-of-the-art models on the ADE20K validation set. Our approach achieves **46. 99%** mIoU score, outperforms the previous state-of-the-art methods by **0. 72%,** which is significant due to the fact that this benchmark is very competitive. CPNet achieves previous best performance among those methods and utilizes the learned context prior with the supervision of the affinity loss to capture the intra-class and inter-class contextual dependencies. In contrast, our FLANet try to capture both spatial-wise and channel-wise dependencies in a single attention map and achieve better performance.

Experiments on the PASCAL VOC Dataset

To verify the generalization of our proposed FLANet, we conduct experiments on the PASCAL VOC dataset. The comparison results are shown in Tab. 5. FLANet based on ResNet101 and HRNetW48 achieves comparable performance on the PASCAL VOC test set, even when other methods are pretrained on additional data.

Conclusions and Future Work

In this paper, we find that traditional self-attention methods suffer from the attention missing problem caused by matrix multiplication. To mitigate this issue, we reformulate the self-attention mechanism into a fully attentional manner

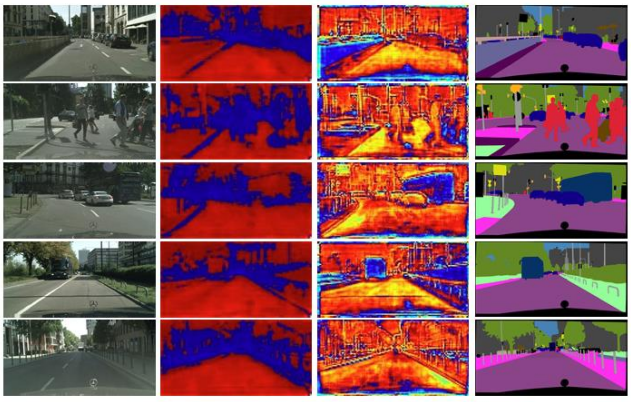


图 6：在 Cityscapes 验证集上，Channel NL 和我们的 FLA 的有人值守特征图的可视化，其中特征图是通过沿通道维度平均来可视化的。

所有图像的远处极点和物体边界均高亮显示。特别是，FLA还可以区分不同的类别，例如，第三排的公共汽车和汽车。这些可视化结果进一步证明了我们提出的模块能够捕获空间相似性并将其编码到频道注意力图中，以实现全注意力。

在ADE20K数据集上的实验

为了进一步验证我们的FLANet的有效性，我们在ADE20K数据集上进行了实验，这是一个具有挑战性的场景解析数据集，包括室内和室外图像。表4报告了FLANet 与ADE20K验证集上最先进模型之间的性能比较。我们的方法达到了 46. 99% 的 mIoU 分数，比以前最先进的方法高出 0. 72%，这很重要，因为这个基准测试非常有竞争力。CPNet 在这些方法中取得了先前的最佳性能，并利用先前学习的上下文来监视亲和力损失来捕获类内和类间的上下文依赖性。相比之下，我们的 FLANet 尝试在单个注意力图中捕获空间和通道上的依赖关系，并获得更好的性能。

PASCAL VOC数据集上的实验

为了验证我们提出的FLANet的泛化性，我们在PASCAL VOC数据集上进行了实验。比较结果如表5所示。基于 ResNet101 和 HRNetW48 的 FLANet 在 PASCAL VOC 测试集上取得了相当的性能，即使其他方法在额外数据上进行了预训练也是如此。

结论和今后的工作

在本文中，我们发现传统的自注意力方法存在由矩阵乘法引起的注意力缺失问题。为了缓解这个问题，我们将自我关注机制重新制定为完全关注的方式

Method	Backbone	mloU
Simple Backbone		
AttaNet (Song, Mei, and Huang 2021)	Res10	43. 71
CCNet (Huang et al. 2019)	Res101	45. 22
ANNet(Zhu et al. 2019)	Res101	45. 24
GFF(Liet al. 2020	Res101	45. 33
OCNet(Yuanet al. 2021)	Res10	45. 40
DMNet(He, Deng, and Qiao 2019)	Res101	45. 50
RecoNet (Chen et al. 2020)	Res101	45. 54
ACNet(Fu et al. 2019b)	Res101	45. 90
DNL(Yin et al. 2020)	Res101	45. 97
CPNet(Yu et al. 2020)	Res101	46. 27
FLANet (Ours)	Resl0	**46. 68**
Advanced Backbone		
HRNetV2(Sun et al. 2019)	HRNetW48	42. 99
DANet (Fu etal. 2019a)	Res101+MG	45. 22
OCNet (Yuan et al. 2021)	HRNetW48	45. 50
DNL(Yin et al. 2020)	HRNetW48	45. 82
FLANet (Ours)	HRNetW48	**46. 99**

Table 4:Comparisons on the ADE20K validation set.

Method	Backbone	mloU
Simple Backbone		
DeepLabv3(Chen et al. 2017)	Res101	85. 7
EncNet (Zhang et al. 2018)	Res101	85. 9
DFN(Yu et al. 2018	Res101	86. 2
CFNet (Zhang, Wang, and Xie 2019)	Res101	87. 2
EMANet (Liet al. 2019)	Res101	87. 7
DeeplabV3+(Chen et al. 2018)	Xception	89. 0
RecoNet (Chen et al. 2020)	Res101	88. 5
FLANet (Ours)	Res101	87. 9
Advanced Backbone		
EMANet (Liet al. 2019)	Res150	88. 2
RecoNet (Chen et al. 2020)	Res150	89. 0
FLANet(Ours)	HRNetW 48	88. 5

Table 5:Comparisons on the PASCALVOC. t indicates that FLANet is trained without using COCO-pretrained model.

which can capture both channel and spatial attentions with a single attention map and also with much less computational complexity. Specifically, we construct global contexts to in- troduce spatial interactions into the channel attention maps. Our FLANet achieves outstanding performance on three se- mantic segmentation datasets. Besides, we also consider the way of introducing channel interactions into the traditional Spatial NL. However, the extremely high computational load limits its practical application. In the future, we will try to achieve that in a more efficient way.

Acknowledgments

This work was supported in part by Shenzhen Natural Sci- ence Foundation under Grant JCYJ20190813170601651, and in part by Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grant AC01202101006 and Grant AC01202101010.

方法	主干网	mloU
简易主干网		
AttaNet (Song, Mei, and Huang 2021)	Res10	43. 71
CCNet (Huang et al. 2019)	Res101	45. 22
ANNet(Zhu et al. 2019)	Res101	45. 24
GFF(Liet al. 2020	Res101	45. 33
OCNet(Yuanet al. 2021)	Res10	45. 40
DMNet(He, Deng, and Qiao 2019)	Res101	45. 50
RecoNet (Chen et al. 2020)	Res101	45. 54
ACNet(Fu et al. 2019b)	Res101	45. 90
DNL(Yin et al. 2020)	Res101	45. 97
CPNet(Yu et al. 2020)	Res101	46. 27
FLANet (Ours)	Resl0	**46. 68**
先进主干网		
HRNetV2(Sun et al. 2019)	HRNetW48	42. 99
DANet (Fu etal. 2019a)	Res101+MG	45. 22
OCNet (Yuan et al. 2021)	HRNetW48	45. 50
DNL(Yin et al. 2020)	HRNetW48	45. 82
FLANet (Ours)	HRNetW48	**46. 99**

表4：ADE20K验证集的比较。

方法	主干网	mloU
简易主干网		
DeepLabv3(Chen et al. 2017)	Res101	85. 7
EncNet (Zhang et al. 2018)	Res101	85. 9
DFN(Yu et al. 2018	Res101	86. 2
CFNet (Zhang, Wang, and Xie 2019)	Res101	87. 2
EMANet (Liet al. 2019)	Res101	87. 7
DeeplabV3+(Chen et al. 2018)	Xception	89. 0
RecoNet (Chen et al. 2020)	Res101	88. 5
FLANet (Ours)	Res101	87. 9
先进主干网		
EMANet (Liet al. 2019)	Res150	88. 2
RecoNet (Chen et al. 2020)	Res150	89. 0
FLANet(Ours)	HRNetW 48	88. 5

表 5：在 PASCAL VOC. t 上的比较表明，FLANet 是在未使用 COCO 预训练模型的情况下进行训练的。

它可以通过单一的注意力图来捕获通道和空间注意力，而且计算复杂度也要低得多。具体来说，我们构建了全球语境，将空间交互引入到频道注意力图中。我们的 FLANet在三个语义分割数据集上取得了出色的性能。此外，我们还考虑了在传统的空间无人学习中引入频道交互的方式。然而，极高的计算负载限制了其实际应用。在未来，我们将尝试以更有效的方式实现这一目标。

鸣谢

这项工作部分得到了深圳自然科学基金会JCYJ20190813170601651资助，部分得到了深圳人工智能与机器人社会研究院AC01202101006资助和AC01202101010资助。

References

Babiloni, F;Marras, I. ;Slabaugh, G. ;and Zafeiriou, S 2020. TESA:Tensor Element Selif-Attention via Matricization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 13942-13951.*

Cao, Y. ;Xu, J. ;Lin, S. ;Wei, F;and Hu, H. 2019. GCNet:Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. *2019 IEEE/CVF International Conference on Computer Vision Workshop(ICCVW), 1971-1980.*

Chen, L. -C. ;Papandreou, G. ;Schroff, F. ;and Adam, H. 2017. Rethinking atrousconvolution for semantic image segmentation. arXiv preprint arXiv:1706. 05587.

Chen, L-C. ;Zhu, Y;Papandreou, G. ;Schroff, F;and Adam, H. 2018. Encoder-decoder withatrous separable convolution for semantic image segmentation. In Proceedings *of the European Conference on Computer Vision(ECCV),* 801-818.

Chen, W. ;Zhu, X. ;Sun, R. ;He, J. ;Li, R. ;Shen, X. ;and Yu, B. 2020. Tensor Low-Rank Reconstruction for Semantic *Segmentation. In Procedings of the European Conference on Computer Vision(ECCv)*

Cheng, B. ;Chen, L. -C. ;Wei, Y. ;Zhu, Y;Huang, Z. ;Xiong J. ;Huang, T. ;Hwu, W. -M. ;and Shi, H. 2019. SPGNet: Semantic Prediction Guidance for Scene Parsing. 2019 *IEEE/CVF International Conference on Computer Vision* (ICCV), 5217-5227.

Choi, S. ;Kim, J. ;and Choo, J. 2020. Cars Can't Fly Up in the Sky:Improving Urban-Scene Segmentation via Height-Driven AttentionNetworks. 2020 IEEE/CVFConferenceon *Computer Vision and Pattern Recognition(CVPR), 9370-*9380.

Chorowski, J. ;Bahdanau, D. ;Serdyuk, D;Cho, K. ;and Bengio, Y. 2015. Attention-Based Models for Speech *Recognition. InAdvances in Neural Information Processing Systems(NeurIPS).*

Cordts, M. ;Omran, M. ;Ramos, S. ;Rehfeld, T;Enzweiler M. ;Benenson, R. ;Franke, U. ;Roth, S. ;and Schiele, B 2016. The cityscapes dataset for semantic urban scene understanding. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 3213-*3223.

Everingham, M. ;Gool, L. ;Williams, C. K. I;Winn, J. ; and Zisserman, A. 2009. The Pascal Visual Object Classes *(VOC)Challenge. International Journal of Computer Vi-*sion, 88:303-338.

Fu, J. ;Liu, J. ;Tian, H;Li, Y. ;Bao, Y. ;Fang, Z. ;and Lu, H. 2019a. Dual attention network for scene segmentation *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3146-3154.*

Fu, J. ;Liu, J. ;Wang, Y. ;Li, Y. ;Bao, Y;Tang, J. ;and Lu, H 2019b. Adaptive context network for scene parsing. In Pro-*ceedings of the IEEEInternationalConferenceon Computer Vision(ICCV), 6748-6757.*

He, J. ;Deng, Z. ;and Qiao, Y. 2019. Dynamic Multi-Scale Filters for Semantic Segmentation. 2019 IEEE/CVF In-*ternational Conference on Computer Vision(ICCV), 3561-*3571.

He, K;Zhang, X. ;Ren, S. ;and Sun, J. 2016. Identity mappings in deep residual networks. In European Conference on Computer Vision (ECCV), 630-645. Springer

Hu, J. ;Shen, L. ;and Sun, G. 2018. Squeze-and-excitation *networks. In Proceedings of the IEEE conferenceon computer vision and pattern recognition, 7132-7141.*

Huang, z. ;Wang, X. ;Huang, L;Huang, C. ;Wei, Y;and Liu, W. 2019. Ccnet:Criss-cross attentionfor semantic segmentation. *In Proceedings of the IEEE International Con. f Ke r, e n-cWe;Hnwa Cnogm, Jp-ut e;Liu V, i s,i o nd (ICCV), 6 03-6* 1 2018. Adaptive Affinity Fields for Semantic Segmentation. In Proceedings *of the European Conference on ComputerVision(ECCV).*

Li, X;Zhang, L. ;You, A. ;Yang, M. ;Yang, K. ;and Tong, Y 2019. Global Aggregation then Local Distribution in Fully Convolutional Networks. In British Machine Vision Confer. *ence (BMVC).*

Li, X. ;Zhao, H;Han, L. ;Tong, Y. ;Tan, S. ;and Yang, K 2020. Gated Fully Fusion for Semantic Segmentation. In *Association for the Advancement of Artificial Intelligence* (AAAI).

Li, X. ;Zhong, Z. ;Wu, J. ;Yang, Y;Lin, Z. ;and Liu, H. 2019. Expectation-Maximization Attention Networks for *Semantic Segmentation. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 9167-9176.*

Long, J. ;Shelhamer, E. ;and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In Proceed *ings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR), 3431-3440*

Ma, N. ;Zhang, X. ;Zheng, H-T. ;and Sun, J. 2018. Shuf fenet v2:Practical guidelines for efficient cnn architecture *design. In Proceedings ofthe European conference on computer vision(ECCV), 116-131.*

Ramachandran, P;Parmar, N. ;Vaswani, A. ;Bello, I;Levskaya, A. ;and Shlens, J. 2019. Stand-Alone Self-Attention *in Vision Models. In Advances in Neural Information Pro. cessing Systems.*

Ronneberger, O. ;Fischer, P;and Brox, T. 2015. U-net:Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention, 234-241. Springer.*

Russakovsky, O. ;Deng, J. ;Su, H. ;Krause, J. ;Satheesh, S. ; Ma, S. ;Huang, Z. ;Karpathy, A. ;Khosla, A;Bernstein, M. ; et al. 2015. Imagenet large scale visual recognition chal*lenge. International journal of computer vision, 115(3):* 211-252.

Song, Q. ;Mei, K. ;and Huang, R. 2021. AttaNet:Attention-Augmented Network for Fast and Accurate Scene Parsing In AAAI.

Sun, K;Xiao, B. ;Liu, D. ;and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Esti-

mation. 2019 IEEE/CVF Conference on Computer Vision and PatternRecognition(CVPR), 5686-5696.

Vaswani, A. ;Shazeer, N. ;Parmar, N. ;Uszkoreit, J. ;Jones L. ;Gomez, A. N. ;Kaiser, L. ;and Polosukhin, I. 2017. At*tention is all you need. In Advances in Neural Information Processing Systems(NeurIPS), 5998-6008.*

Wang, X. ;Girshick, R. ;Gupta, A. ;and He, K. 2018. Non-*local neural networks. In Proceedings of the IEEE Confer- ence on Computer Vision and Pattern Recognition(CVPR),* 7794-7803.

Yang, M. ;Yu, K. ;Zhang, C. ;Li, Z. ;and Yang, K. 2018. DenseASPP for Semantic Segmentation in Street Scenes. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 3684-3692.*

Yin, M. ;Yao, Z. ;Cao, Y. ;Li, X. ;Zhang, Z. ;Lin, S. ;and Hu, H. 2020. Disentangled Non-Local Neural Networks *In Proceedings of the European Conference on Computer Vsion(ECCV).*

Yu, C;Wang, J. ;Gao, C. ;Yu, G. ;Shen, C. ;and Sang, N. 2020. Context Prior for Scene Segmentation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 12413-12422.*

Yu, C. ;Wang, J. ;Peng, C. ;Gao, C. ;Yu, G. ;and Sang, N. 2018. Learning a discriminative feature network for seman-*tic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR), 1857-* 1866.

Yuan, Y;Chen, X. ;and Wang, J. 2020. Object-Contextual Representations for Semantic Segmentation. In Proceedings *of the European Conferenceon Computer Vision(ECCV)*

Yuan, Y. ;Huang, L. ;Guo, J. ;Zhang, C. ;Chen, X. ;and Wang, J. 2021. OCNet:Object Context for Semantic Segmentation. Int. J. Comput. Vis., 129:2375-2398.

Zhang, F;Chen, Y. ;Li, Z. ;Hong, Z. ;Liu, J. ;Ma, F. ;Han, J. ;and Ding E. 2019a Acfnet:Attentional class feature n etwork for › emant ic .segmentation In Proceedings of the *IEEE International Conference on Computer Vision, 6798-* 6807.

Zhang, H. ;Dana, K. ;Shi, J. ;Zhang, Z. ;Wang, X. ;Tyagi, A. ;and Agrawal, A. 2018. Context Encoding for Semantic *Segmentation. 2018 IEEEICVF Conferenceon Computer Vsion and Pattern Recognition(CVPR), 7151-7160.*

Zhang, H. ;Wang, C. ;and Xie, J. 2019. Co-Occurrent Features in Semantic Segmentation. 2019 IEEE/CVF Confer-*ence on Computer Vision and Pattern Recognition(CVPR),* 548-557.

Zhang, L. ;Li, X. ;Arnab, A. ;Yang, K. ;Tong, Y. ;and Torr P. 2019b. Dual Graph Convolutional Network for Semantic Segmentation. In British Machine Vision Conference *(BMVC).*

Zhao, H. ;Shi, J. ;Qi, X. ;Wang, X. ;and Jia, J. 2017. Pyra-*mid scene parsing network. In Proceedings ofthe IEEEcon- ference on computer vision and pattern recognition, 2881-* 2890.

Zhao, H. ;Zhang, Y. ;Liu, S. ;Shi, J. ;Loy, C. C. ;Lin, D. ;and Jia, J. 2018. PSANet:Point-wise Spatial Attention Network *for Scene Parsing. In Proceedings of the European Confer- ence on Computer Vision(ECCV), 270-286.*

Zhou, B. ;Zhao, H. ;Puig, X. ;Fidler, S. ;Barriuso, A. ; and Torralba, A. 2017. Scene Parsing Through ADE20K *Dataset. In Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition(CVPR).*

Zhu, Z. ;Xu, M. ;Bai, S. ;Huang, T. ;and Bai, X. 2019. Asymmetric Non-Local Neural Networks for Semantic Segmentation. *2019 IEEE/CVF International Conference on Computer Vision(ICCV), 593-602.*