# Global Context Networks

# 全局上下文网络

Yue Cao\*, Jiarui Xu\*, Stephen Lin, Fangyun Wei, Han Hu

蔡跃\*，徐佳锐\*，林斯蒂芬，魏方云，胡汉

Abstract—The Non-Local Network (NLNet) presents a pioneering approach for capturing long-range dependencies within an image, via aggregating query-specific global context to each query position. However, through a rigorous empirical analysis, we have found that the global contexts modeled by the non-local network are almost the same for different query positions. In this paper, we take advantage of this finding to create a simplified network based on a query-independent formulation, which maintains the accuracy of NLNet but with significantly less computation. We further replace the one-layer transformation function of the non-local block by a two-layer bottleneck, which further reduces the parameter number considerably. The resulting network element, called the global context (GC) block, effectively models global context in a lightweight manner, allowing it to be applied at multiple layers of a backbone network to form a global context network (GCNet). Experiments show that GCNet generally outperforms NLNet on major benchmarks for various recognition tasks. The code and network configurations are available at https://github.com/xvjiarui/GCNet

摘要—非局部网络（NLNet）提出了一种通过聚合特定查询的全局上下文到每个查询位置来捕获图像中长距离依赖的开创性方法。然而，通过对非局部网络建模的全局上下文进行严格的实证分析，我们发现不同查询位置的全局上下文几乎相同。在本文中，我们利用这一发现构建了一个基于查询无关公式的简化网络，该网络保持了NLNet的准确性，但计算量大大减少。我们进一步将非局部块中的一层变换函数替换为两层瓶颈，这显著减少了参数数量。由此产生的网络元素，称为全局上下文（GC）块，能够以轻量级的方式有效地建模全局上下文，允许它在主干网络的多个层上应用，形成一个全局上下文网络（GCNet）。实验表明，GCNet在各种识别任务的主要基准上通常优于NLNet。代码和网络配置可在 https://github.com/xvjiarui/GCNet 获取。

Index Terms-deep network, self-attention model, global context, object detection.

关键词—深度网络，自注意力模型，全局上下文，目标检测。

# 1 Introduction

# 1 引言

Long-range dependencies among pixels in an image are essential to capture for global understanding of a visual scene. This dependency modeling is proven to benefit a wide range of recognition tasks, such as image classification [2], object detection and segmentation [3], [4], and video action recognition [5]. In convolutional neural networks, long-range dependencies are mainly modeled by deep stacking of convolution layers, where each layer models pixel relationships within a local neighborhood. However, direct repetition of convolution layers is computationally inefficient and hard to optimize [5], due in part to difficulties in delivering messages between distant positions.

图像中像素之间的长距离依赖性对于视觉场景的全局理解至关重要。这种依赖性建模已被证明可以受益于广泛的识别任务，如图像分类[2]、目标检测和分割[3]、[4]，以及视频行为识别[5]。在卷积神经网络中，长距离依赖性主要通过深度堆叠卷积层来建模，其中每一层都建模了局部邻域内的像素关系。然而，直接重复卷积层在计算上是低效的，并且难以优化[5]，部分原因是远距离位置之间的消息传递困难。

To address this issue, the non-local network (NLNet) [5] utilizes a layer to model long-range dependencies, via a self-attention mechanism [6]. For each query position, the non-local network first computes pairwise relations between the query position and all other positions to form an attention map, and then aggregates the features of all positions by a weighted sum with the weights defined by the attention map. The aggregated features are finally added to the features of each query position to form the output.

为了解决这个问题，非局部网络（NLNet）[5] 使用一个层来通过自注意力机制 [6] 模拟长距离依赖。对于每个查询位置，非局部网络首先计算查询位置与所有其他位置之间的成对关系，形成一个注意力图，然后通过注意力图定义的权重，将所有位置的特征进行加权求和。最后，将聚合后的特征加到每个查询位置的特征上，形成输出。

The query-specific attention weights in the non-local network are expected to reflect the importance of the corresponding positions to the query position. Visualizing these weights would help to better understand their behavior, but such analysis was largely missing in the original paper. In an analysis that we conducted, a surprising observation can be made. As shown in Figure 1 we found that the attention maps for different query positions are almost the same, indicating that the learnt dependency is basically query-independent. This observation is further verified by the statistical analysis in Tables 1, 2 and 3 , which show that the distance between the attention maps of different query positions is very small. This observation is verified in three standard tasks, object detection on COCO, image classification on ImageNet and action recognition on Kinetics.

在非局部网络中，查询特定的注意力权重预计会反映相应位置对查询位置的重要性。可视化这些权重将有助于更好地理解它们的行为，但原始论文中很大程度上缺失了此类分析。在我们进行的一项分析中，可以观察到令人惊讶的现象。如图1所示，我们发现不同查询位置的注意力图几乎相同，表明学到的依赖基本上是查询无关的。这一观察进一步得到了表1、2和3的统计分析的验证，这些表显示不同查询位置的注意力图之间的距离非常小。这一观察在三个标准任务中得到了验证，分别是COCO对象检测、ImageNet图像分类和Kinetics动作识别。

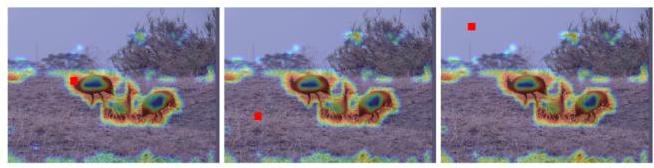


Fig. 1: Visualization of attention maps (heatmaps) for different query positions (red points) in a non-local block on COCO object detection. The three attention maps are all almost the same. More examples are presented in Figure 2

图1：在COCO对象检测的非局部块中，不同查询位置（红点）的注意力图（热图）可视化。这三个注意力图几乎都相同。更多示例展示在图2中。

Based on this observation, we propose a simplification of the non-local block in which a query-independent attention map is explicitly used for all query positions. The output is then formed by the same aggregation of features using this attention map as weights. This simplified block requires significantly less computation than the original non-local block, but exhibits almost no decrease in accuracy on several important visual recognition tasks. The block design follows a general three-step framework: (a) a context modeling module which aggregates the features of all positions together to form a global context feature; (b) a feature transform module to capture the channel-wise interdependencies; and (c) a fusion module to merge the global context feature into features of all positions. We further significantly reduce the parameter number by replacing the one-layer transformation function of the non-local block with a bottleneck of two layers, to form a new unit that we call the global context (GC) block.

基于这一观察，我们提出了对非局部块的简化，其中查询无关的注意力图被显式地用于所有查询位置。然后输出通过使用该注意力图作为权重对特征进行相同的聚合形成。这个简化的块比原始的非局部块计算量大大减少，但在几个重要的视觉识别任务上几乎没有任何准确度下降。块的设计遵循一个通用的三步框架：（a）一个上下文建模模块，它将所有位置的特征聚合在一起形成全局上下文特征；（b）一个特征转换模块，用于捕获通道间的相互依赖；以及（c）一个融合模块，将全局上下文特征合并到所有位置的特征中。我们还通过将非局部块的单层转换函数替换为两层瓶颈结构，进一步显著减少了参数数量，从而形成我们称之为全局上下文（GC）块的新单元。

Because of the lightweight computation of the GC block, it can be applied to all residual blocks in the ResNet architecture, in contrast to the original non-local block which is usually applied after just one or a few layers due to its heavy processing. We refer to this network as the global context network (GCNet). On COCO object detection/instance segmentation, it is found that GCNet outperforms NLNet by on and on with just a relative increase in FLOPs. In addition, GCNet yields significant performance gains over four general visual recognition tasks: object detection/segmentation on COCO on APbbox, and on over Mask R-CNN with FPN and ResNet-50 as backbone [7]), semantic segmentation on Cityscapes (3.2% on mIoU over ResNet-101 as backbone with dilated convolutions), image classification on ImageNet on top-1 accuracy over ResNet-50 [8]), and action recognition on Kinetics on top- 1 accuracy over the ResNet-50 Slow-only baseline [9]), with less than a increase in computation cost.

由于 GC 块的计算轻量，它可以应用于 ResNet 架构中的所有残差块，而原始的非局部块由于其处理繁重，通常仅在一层或几层之后应用。我们称这个网络为全局上下文网络（GCNet）。在 COCO 目标检测/实例分割任务上，发现 GCNet 在 上比 NLNet 性能高出 ，在 上高出 ，仅增加了 的 FLOPs。此外，GCNet 在四个通用视觉识别任务上带来了显著的性能提升：在 COCO 上的目标检测/分割任务中 在 APbbox 上，以及在 上超过带有 FPN 和 ResNet-50 作为主干的 Mask R-CNN [7]；在 Cityscapes 上的语义分割任务中 在 mIoU 上超过带有扩张卷积的 ResNet-101 作为主干；在 ImageNet 上的图像分类任务中 在 top-1 准确率上超过 ResNet-50 [8]；以及在 Kinetics 上的动作识别任务中 在 top-1 准确率上超过 ResNet-50 Slow-only 基线 [9]，计算成本的增加不到 。

* \* Equal contribution.
* \* 相同贡献。
* Yue Cao, Stephen Lin, Fangyun Wei and Han Hu are with Microsoft Research Asia. Jiarui Xu is with Hong Kong University of Science and Technology. This work was done when Jiarui Xu was an intern at Mi-crosoft Research Asia. E-mail: Yue Cao (yuecao@microsoft.com), Jiarui Xu (xvjiarui0826@gmail.com), Stephen Lin (stevelin@microsoft.com), Fangyun Wei (fawe@microsoft.com). Correspondence to: Han Hu (hanhu@microsoft.com).
* 赵曹（Yue Cao）、林斯蒂芬（Stephen Lin）、魏方云（Fangyun Wei）和胡汉（Han Hu）在微软亚洲研究院工作。徐嘉锐（Jiarui Xu）在香港科技大学工作。这项工作是在徐嘉锐在微软亚洲研究院实习期间完成的。电子邮箱：赵曹（yuecao@microsoft.com），徐嘉锐（xvjiarui0826@gmail.com），林斯蒂芬（stevelin@microsoft.com），魏方云（fawe@microsoft.com）。请联系：胡汉（hanhu@microsoft.com）。
* A preliminary version of this manuscript was published in [1].
* 本文的初步版本已发表于 [1]。

# 2 Related Work

# 2 相关工作

# 2.1 Deep architectures

# 2.1 深度架构

Recent progress in computer vision have largely been driven by the improvement of basic deep architectures, which extract features for visual elements. One direction of improvement is to design better functional formulations of basic components to elevate the power of deep networks for the general purpose of image feature extraction. A pioneering work along this path is AlexNet [10], which proves that increasing the depth and width of convolutional neural networks can achieve impressive accuracy in classifying objects in ImageNet. Since then, vast improvements have been made to unleash the power of deep architectures. VGG [11] further increases the depth and width, and replaces most large-kernel convolution layers by smaller ones of , which has become widely used in subsequent architecture designs. GoogLeNet [12] extends the idea of the multi-branch layer from NIN [13] and introduces convolution to reduce the number of parameters. ResNet [8] introduces skip connections (also called shortcuts), which can significantly reduce the gradient vanishing issue and allows the network to be tremendously deep. In DenseNet [14], every layer obtains additional inputs from all preceding layers and passes its own feature maps to all subsequent layers, through concatenation operations. ResNeXt [15] and Xception [16] adopt group convolution to increase cardinality and reduce the redundancy of the network parameters. Deformable Convolution Networks [17], [18] present deformable convolutions for enhancing geometric modeling ability, which can significantly improve performance on fine-grained recognition tasks. Local Relation Networks [19] replace all spatial convolution layers by local relation layers, which adaptively determine aggregation weights based on the compositional relationship of local pixel pairs. Different from handcrafted architectures, automatic search of the cell structure for deep architectures has attracted much attention recently [20], [21].

计算机视觉领域的近期进展在很大程度上是由基本深度架构的改进推动的，这些架构用于提取视觉元素的特征。改进的一个方向是设计基本组件的更好功能公式，以提高深度网络在图像特征提取通用目的上的能力。沿着这条路径的开创性工作是AlexNet [10]，它证明了增加卷积神经网络的深度和宽度可以在ImageNet中对对象分类达到令人印象深刻的准确性。自那时起，为了发挥深度架构的力量，已经取得了巨大的改进。VGG [11]进一步增加了深度和宽度，并将大多数大核卷积层替换为 的小核层，这已成为后续架构设计中广泛使用的方法。GoogLeNet [12]扩展了NIN [13]的多分支层思想，并引入了 卷积以减少参数数量。ResNet [8]引入了跳接（也称为捷径），可以显著减少梯度消失问题，并允许网络非常深。在DenseNet [14]中，每一层都会从所有先前层获得额外的输入，并通过连接操作将其特征图传递给所有后续层。ResNeXt [15]和Xception [16]采用组卷积来增加基数并减少网络参数的冗余。可变形卷积网络[17]、[18]提出了可变形卷积以增强几何建模能力，这在细粒度识别任务上可以显著提高性能。局部关系网络[19]用局部关系层替换了所有空间卷积层，这些层基于局部像素对的组合关系自适应地确定聚合权重。与手工制作的架构不同，最近深度架构的单元结构自动搜索已经引起了广泛关注[20]、[21]。

Another direction of improvement is to invent deep architectures for specific tasks, such as semantic segmentation [22], [23], [24], [25], [26], [27], object detection [3], [28], [29], [30], [31], and video action recognition [9], [32], [33], [34]. MobileNet [35], [36] is designed to adopt depthwise separable convolution as the basic block for mobile and embedded vision applications. Shuf-fleNet [37], [38] adopts channel shuffling, which facilitates the use of group convolution with convolutions. Fully-convolutional Networks (FCN) [22], [24], [26] are designed to make dense predictions for per-pixel tasks like semantic segmentation. The YOLO series [29], [30] frames object detection as the regression of spatially separated bounding boxes and associated class probabilities, which is both fast and effective. For video action recognition tasks, to better incorporate temporal information in feature extraction, I3D [39] introduces 3D convolution to deep networks. To reduce the computation cost, P3D [40] separates the 3D convolution into a sequence of temporal-only convolution and spatial-only convolution.

另一个改进方向是为特定任务发明深度架构，例如语义分割 [22]、[23]、[24]、[25]、[26]、[27]、目标检测 [3]、[28]、[29]、[30]、[31] 以及视频行为识别 [9]、[32]、[33]、[34]。MobileNet [35]、[36] 被设计为采用深度可分离卷积作为移动和嵌入式视觉应用的基本块。ShuffleNet [37]、[38] 采用通道混洗，这有助于使用带有 卷积的组卷积。全卷积网络（FCN）[22]、[24]、[26] 被设计用于为每个像素的任务如语义分割进行密集预测。YOLO 系列算法 [29]、[30] 将目标检测框架化为空间分离的边界框及其相关类概率的回归，这种方法既快速又有效。对于视频行为识别任务，为了在特征提取中更好地融入时间信息，I3D [39] 在深度网络中引入了3D卷积。为了降低计算成本，P3D [40] 将3D卷积分解为仅时间卷积和仅空间卷积的序列。

The proposed global context network is a new architecture designed for general purpose. It introduces a novel global context block which models long-range information into existing architectures, showing general improvements on a wide range of vision tasks, such as object detection, instance segmentation, image classification and action recognition.

提出的全局上下文网络是一种为通用目的而设计的新架构。它引入了一种新颖的全局上下文块，将长距离信息建模到现有架构中，并在广泛的视觉任务上显示出普遍的改进，例如目标检测、实例分割、图像分类和行为识别。

# 2.2 Long-range dependency modeling

# 2.2 长距离依赖建模

While existing deep architectures mainly work by stacking layers which operate locally, there are also methods that directly model long-range dependency using a single layer. Such methods can be categorized into two classes: pairwise based, and context fusion based.

虽然现有的深度架构主要通过堆叠局部操作的层来工作，但也有方法直接使用单一层建模长距离依赖。这类方法可以分为两类：基于成对的方法和基于上下文融合的方法。

Most pairwise methods are based on the self-attention mechanism, and the non-local network (NLNet) is a pioneering work [5] for pixel-pixel pairwise relation modeling that has proven beneficial for several visual recognition tasks, such as object detection and action recognition. There are also extensions of non-local networks proposed to benefit specific tasks. Object Context Networks (OCNet) [41] model pixel-wise relationships in the same object category via self-attention mechanisms and also capture context at multiple scales. Dual Attention Networks (DANet) [42] use self-attention mechanisms to model pixel-pixel relationships and channel-channel relationships to improve feature representations. Criss-Cross Networks (CCNet) [43] accelerate NLNet via stacking two criss-cross blocks, which can enlarge the dependency range to the whole feature map with low computational cost.

大多数成对方法基于自注意力机制，而非局部网络（NLNet）是像素间成对关系建模的开创性工作 [5]，已在多个视觉识别任务中证明其益处，例如目标检测和动作识别。还提出了非局部网络的扩展以利于特定任务。对象上下文网络（OCNet）[41] 通过自注意力机制建模同一对象类别中的像素级关系，并捕获多尺度下的上下文。双重注意力网络（DANet）[42] 使用自注意力机制来建模像素间关系和通道间关系，以改进特征表示。十字交叉网络（CCNet）[43] 通过堆叠两个十字交叉块加速 NLNet，这可以以低计算成本将依赖范围扩展到整个特征图。

While it is widely believed that NLNet benefits visual recognition due to pairwise relation modeling, this paper empirically proves that such belief is actually incorrect. In fact, for several important visual recognition tasks such as ImageNet image classification, COCO object detection and Kinetics action recognition, we observe that NLNet degenerates to learning the same global context vector for different pixels, and thus the effectiveness of NLNet can mainly be ascribed to global context modeling other than pairwise relation modeling. For some other visual recognition tasks, such as semantic segmentation, although we observe that some kind of pairwise relation is learnt, the accuracy improvement is still mostly ascribed to its global context modeling ability. Based on this observation, we propose a simplification of the non-local block, which explicitly learns global context other than pairwise relations. The resulting block, called the global context (GC) block, consumes significantly less computation than the non-local block but performs with the same accuracy on several important tasks. Note while the proposed GC block exploits the findings of this degeneration issue to explicitly simplify the non-local block, in a follow-up to this paper, our work on disentangled non-local networks (DNL) [44] on the contrary attempts to alleviate this degeneration problem by a disentangled design in a manner that allows learning of different contexts for different pixels while preserving the shared global context.

虽然人们普遍认为 NLNet 由于成对关系建模而有利于视觉识别，但本文通过实证证明这种观点实际上是不正确的。实际上，对于像 ImageNet 图像分类、COCO 目标检测和 Kinetics 动作识别等几个重要的视觉识别任务，我们观察到 NLNet 退化为学习不同像素的相同全局上下文向量，因此 NLNet 的有效性主要可以归因于全局上下文建模，而不是成对关系建模。对于其他一些视觉识别任务，如语义分割，尽管我们观察到学到了某种成对关系，但准确性的提高仍然主要归因于其全局上下文建模能力。基于这一观察，我们提出了非局部块的一个简化版本，该版本明确地学习全局上下文而不是成对关系。这个结果块，称为全局上下文（GC）块，在多个重要任务上的计算消耗比非局部块显著减少，但性能相同。注意，尽管提出的 GC 块利用了这种退化解体问题的发现来明确简化非局部块，但在本文的后续工作中，我们对解耦非局部网络（DNL）[44] 的工作则相反，试图通过解耦设计来缓解这种退化解体问题，该设计允许为不同像素学习不同的上下文，同时保留共享的全局上下文。

Different from pairwise methods, context fusion methods operate by strengthening the feature of each position by a context feature that aggregates information from all pixels including those at long range. For example, SENet fuse the two features by adaptive rescaling on different channels. GENet [45] uses local patches to compute position-adaptive context features. PSANet [46] proposes to connect each position on the feature map to all the other ones through a self-adaptively learned attention mask, and aggregate the features of other positions via rescaling. CBAM [47] recalibrates the importance of both different spatial positions and channels also via rescaling. All these methods adopt rescaling for feature aggregation, which may be of limited effectiveness for global context modeling.

与成对方法不同，上下文融合方法通过增强每个位置的特征，使用一个聚合了所有像素（包括远距离像素）信息的上下文特征。例如，SENet 通过在不同通道上进行自适应缩放来融合两个特征。GENet [45] 使用局部补丁来计算位置自适应的上下文特征。PSANet [46] 提出通过一个自学习自适应的注意力掩码将特征图上的每个位置与其他所有位置连接起来，并通过缩放来聚合其他位置的特征。CBAM [47] 也通过缩放重新校准不同空间位置和通道的重要性。所有这些方法都采用缩放进行特征聚合，这对于全局上下文建模可能效果有限。

The proposed GCNet is also a context fusion method. But by using a different context feature computation method (attention pooling) and a different fusion method (addition), GCNet performs generally better than the widely used SENet method. Noting that the context feature computation and fusion methods used in GCNet are inherited from NLNet, the proposed GCNet can be also seen as a product of connecting two representative long-range dependency modeling methods, NLNet and SENet, but makes good use of their respective strengths (GCNet is the same as NLNet in better context modeling and information fusion, while being as lightweight as SENet).

所提出的GCNet也是一种上下文融合方法。但是通过使用不同的上下文特征计算方法（注意力池化）和不同的融合方法（加法），GCNet通常比广泛使用的SENet方法表现更好。注意到GCNet中使用的上下文特征计算和融合方法继承自NLNet，所提出的GCNet也可以看作是连接两种代表性的长距离依赖建模方法NLNet和SENet的产物，但很好地发挥了它们各自的优势（GCNet在上下文建模和信息融合方面与NLNet相同，同时轻量级如SENet）。

# 2.3 Self-attention modeling

# 2.3 自注意力建模

This paper is also related to the general self-attention mechanism whose application extends beyond pixel relation modeling [3], [5], [6], [19], [41], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60].

本文还与一般性的自注意力机制相关，其应用范围超出了像素关系建模 [3], [5], [6], [19], [41], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60]。

In natural language processing, Transformer [6], which applies a self-attention mechanism to model long-range dependencies between words, is a milestone work for machine translation. Graph Attention Networks (GAT) [56] improve graph convolution with self-attention mechanisms that operate on graph-structured data, producing remarkable gains over baseline graph convolution methods. Self-attention Generative Adversarial Networks (SAGAN) [57] generate high-resolution details as a function of not only spatially local points but also distant points, via self-attention mechanisms that model long-range dependency.

在自然语言处理中，Transformer [6] 通过应用自注意力机制来建模单词之间的长距离依赖关系，是机器翻译领域的里程碑之作。图注意力网络（GAT）[56] 通过在图结构数据上应用自注意力机制改进了图卷积，相较于基线图卷积方法取得了显著提升。自注意力生成对抗网络（SAGAN）[57] 通过自注意力机制建模长距离依赖，不仅能够生成空间局部点的细节，还能生成远距离点的细节。

For visual recognition, aside from pixel relation modeling, the attention mechanism is also applied for object-object/object-pixel relation modeling [3], [61], which is proven effective in object detection.

在视觉识别领域，除了像素关系建模之外，注意力机制也应用于对象-对象/对象-像素关系建模 [3], [61]，这已在目标检测中被证明是有效的。

The presented analysis and proposed GCNet in this paper are basically about the general self-attention mechanism, with experiments and instantiations mainly targeting the problem of pixel-pixel relation modeling. Such an analysis and global context modeling approach could be extended to other self-attention applications such as object-object/object-pixel relation modeling, natural language processing, and graph social networks. For these applications, there are questions of whether the pairwise relations can be well learnt by the self-attention mechanism and how the global context modeling approach can effectively contribute. Both of these questions on broader applications are promising directions for further study.

本文提出的分析和提出的GCNet基本上是关于通用自注意力机制的，实验和实例主要针对像素-像素关系建模问题。这种分析和全局上下文建模方法可以扩展到其他自注意力应用，例如对象-对象/对象-像素关系建模、自然语言处理和图社交网络。对于这些应用，存在自注意力机制是否能够很好地学习成对关系以及全局上下文建模方法如何有效贡献的问题。这两个关于更广泛应用的问题都是进一步研究的很有前景的方向。

# 3 Analysis of Non-Local Networks

# 3 非局部网络分析

In this section, we first review the design of the non-local block [5]. While in-depth studies have been rare on what a non-local block learns and what makes it effective, we conduct such a study both qualitatively and statistically. Qualitatively, we visualize the attention maps across different query positions generated by a widely-used instantiation of the non-local block. Statistically, we compute the average cosine distances between different feature maps (including input, attention map, output and so on) inside the non-local block, to delve deep into the non-local block design. This in-depth study brings a new understanding of the non-local block and may inspire new approaches as in the next section.

在本节中，我们首先回顾了非局部块[5]的设计。虽然对非局部块学习的内容以及使其有效的原因的深入研究较少，但我们从定性和统计两个方面进行了此类研究。定性地，我们可视化了一个广泛使用的非局部块实例在不同查询位置生成的注意力图。统计上，我们计算了非局部块内部不同特征图（包括输入、注意力图、输出等）之间的平均余弦距离，以深入探讨非局部块的设计。这种深入研究带来了对非局部块的新理解，并可能激发下一节中提到的新方法。

# 3.1 Revisiting the Non-local Block

# 3.1 重新审视非局部块

The basic non-local block [5] aims at strengthening the features of the query position via aggregating information from other positions. We denote as the feature map of one input instance (e.g., an image or video), where is the number of positions in the feature map (e.g., for image, for video). and denote the input and output of the non-local block, respectively, which have the same dimensions. The nonlocal block is formulated as

基本的非局部块[5]旨在通过从其他位置聚合信息来增强查询位置的特征。我们用 表示一个输入实例（例如，图像或视频）的特征图，其中 是特征图中的位置数（例如，图像的 ，视频的 ）。 和 分别表示非局部块的输入和输出，它们的尺寸相同。非局部块被公式化为

where is the index of query positions, and enumerates all possible positions. denotes the relationship between position and , and has a normalization factor and denote linear transform matrices (e.g.,1x1 convolution). For simplification, we denote as the normalized pairwise relationship between position and .

其中 是查询位置的索引， 枚举所有可能的位置。 表示位置 和 之间的关系，并具有归一化因子 。 和 表示线性变换矩阵（例如，1x1卷积）。为了简化，我们将 表示为位置 和 之间的归一化成对关系。

In [5], four instantiations of the non-local block are provided by defining as different functions:

在[5]中，通过将 定义为不同的函数，提供了非局部块的四种实例化。

* Gaussian. in is the Gaussian function, defined as
* 高斯函数 在 中定义为
* Embedded Gaussian. It is a simple extension of Gaussian by using an embedding space to compute similarity, defined as ;
* 嵌入高斯函数。它通过使用嵌入空间来计算相似度，对高斯函数的简单扩展，定义为 ;
* Dot product. in is defined as a dot-product similarity, formulated as ;
* 点积 在 中定义为点积相似度，公式表示为 ;
* Concat. It is defined as .
* 连接操作。它定义为 。

We illustrate the architecture of two most widely-used instantiations, Embedded Gaussian and Gaussian, in Figure 3(a) and 3(b).

我们在图3(a)和3(b)中展示了两种最广泛使用的实例化架构，嵌入式高斯和普通高斯。

The non-local block can be regarded as a query-specific global context modeling block, which strengthens the feature at a query position by a query-specific global context vector, computed by a weighted sum over all positions. The weights are determined by a similarity between two positions, and the weights over all positions form an attention map for one query position. The time and space complexity of the non-local block are heavy in that they are both quadratic to the number of positions . Likely as a result, it is applied to only a few places in a network architecture, e.g. as one block inserted into the Mask R-CNN framework.

非局部块可以看作是一个查询特定的全局上下文建模块，它通过查询特定的全局上下文向量加强查询位置的特征，该向量是通过所有位置上的加权和计算的。权重由两个位置之间的相似度决定，所有位置上的权重构成一个针对单个查询位置的注意力图。非局部块的时间和空间复杂度较高，因为它们都是位置数量的二次方 。可能正因为如此，它只被应用于网络架构中的少数位置，例如作为插入到Mask R-CNN框架中的一个块。

The non-local block [5] is proven to benefit many visual recognition tasks, such as object detection/instance segmentation, and action recognition. It is believed that such effectiveness arises from effective learning of pairwise pixel relations [5]. Nevertheless, direct evidence and an in-depth study of this has been lacking. In the following, we analyze what is truly learnt in non-local networks, both qualitatively and statistically. Such a study shed light on the behavior of non-local networks.

非局部块 [5] 被证明能够提升许多视觉识别任务，如目标检测/实例分割和动作识别。人们认为这种有效性来源于对像素对关系的有效学习 [5]。然而，关于这一点的直接证据和深入研究一直缺失。在接下来的内容中，我们将定性和统计地分析非局部网络中真正学到了什么。这项研究揭示了非局部网络的行为。

# 3.2 Analysis

# 3.2 分析

# 3.2.1 Visualization

# 3.2.1 可视化

To intuitively understand the behavior of the non-local block, we first visualize the attention maps for different query positions. As different instantiations achieve comparable performance [5], here we only visualize the most widely-used version, Embedded Gaussian, which has the same formulation as the block proposed in [6]. Since attention maps in videos are hard to visualize and understand, we only show visualizations on object detection/instance segmentation, which takes images as input. Following the standard setting of non-local networks for object detection [5], we conduct experiments on Mask R-CNN with FPN and ResNet50, and only add one non-local block right before the last residual block of .

为了直观地理解非局部块的行为，我们首先可视化不同查询位置的关注图。由于不同的实例化方法取得了相当的性能 [5]，在这里我们只可视化最广泛使用的版本，嵌入式高斯，其公式与 [6] 中提出的块相同。由于视频中的关注图难以可视化和理解，我们只展示在对象检测/实例分割上的可视化，该输入为图像。遵循对象检测中非局部网络的标准设置 [5]，我们在带有FPN和ResNet50的Mask R-CNN上进行实验，并且只在 的最后一个残差块之前添加一个非局部块。



Fig. 2: Visualization of attention maps (heatmaps) for different query positions (red points) in a non-local block on COCO object detection. For the same image, the attention maps of different query points are almost the same. Best viewed in color.

图 2：在不同查询位置（红色点）的非局部块上对 COCO 对象检测的可视化关注图（热图）。对于同一图像，不同查询点的关注图几乎相同。彩色查看最佳。

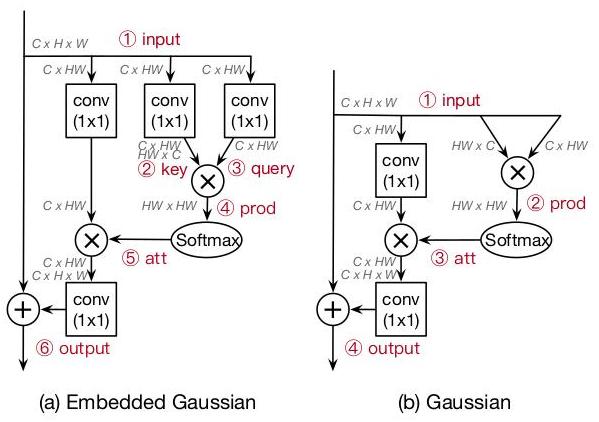


Fig. 3: Two instantiations of the non-local block: Embedded Gaussian and Gaussian. The feature maps are shown by their dimensions, e.g. CxHxW. denotes matrix multiplication, and is broadcast element-wise addition. For two matrices with different dimensions, broadcast operations first broadcast features in each dimension to match the dimensions of the two matrices. The feature maps marked in red (e.g. ’ (5) att’) are statistically analyzed in Tables 1, 3 and 2

图 3：非局部块的两种实例化：嵌入式高斯和高斯。特征图通过它们的尺寸显示，例如 CxHxW。 表示矩阵乘法， 是逐元素加法的广播。对于两个尺寸不同的矩阵，广播操作首先在每个维度上广播特征以匹配两个矩阵的尺寸。在红色标记的特征图（例如 ’ (5) att’）在表 1、3 和 2 中进行了统计分析。

In Figure 2, we randomly select six images from the COCO dataset, and visualize three different query positions (red points) and their query-specific attention maps (heatmaps) for each image. We surprisingly find that for different query positions, their attention maps are almost the same. This suggests that it may be redundant for the non-local block to compute different attention maps for different positions in object detection, as the non-local block may not learn pixel-pixel relationships in this task but rather just global context. This observation motivates us to delve deep into the design of non-local block, to understand its real behavior.

在图2中，我们从COCO数据集中随机选取了六张图像，并可视化了每张图像上三种不同的查询位置（红色点）及其查询特定的注意力图（热图）。令人惊讶的是，我们发现对于不同的查询位置，它们的注意力图几乎相同。这表明，在目标检测任务中，非局部块为不同位置计算不同的注意力图可能是多余的，因为非局部块可能并没有在这个任务中学习像素间的关联，而仅仅是全局上下文。这一观察结果激发我们深入研究非局部块的设计，以了解其真实行为。

| Dataset | Method | APbbox |  | cosine distance | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | input | output | att |
| COCO | baseline | 37.2 | 33.8 |  |  |  |
| Gaussian | 38.0 | 34.8 | 0.397 | 0.062 | 0.177 |
| E-Gaussian | 38.0 | 34.7 | 0.402 | 0.012 | 0.020 |
| Dot product | 38.1 | 34.8 | 0.405 | 0.020 | 0.015 |
| Concat | 38.0 | 34.9 | 0.393 | 0.003 | 0.004 |
| Dataset | Method | Top-1 | Top-5 | input | output | att |
| Kinetics | baseline | 74.9 | 91.9 | 。 |  | - |
| Gaussian | 76.0 | 92.3 | 0.345 | 0.056 | 0.056 |
| E-Gaussian | 75.9 | 92.2 | 0.358 | 0.003 | 0.004 |
| Dot product | 76.0 | 92.3 | 0.353 | 0.095 | 0.099 |
| Concat | 75.4 | 92.2 | 0.354 | 0.048 | 0.049 |
| Dataset | Method | Top-1 | Top-5 | input | output | att |
| ImageNet | baseline | 76.5 | 93.4 | - | - | - |
| Gaussian | 77.1 | 93.6 | 0.045 | 0.005 | 0.011 |
| E-Gaussian | 77.2 | 91.9 | 0.301 | 0.074 | 0.115 |
| Dot product | 77.0 | 93.5 | 0.396 | 0.081 | 0.098 |
| Concat | 76.9 | 93.5 | 0.379 | 0.023 | 0.090 |

TABLE 1: Statistical analysis on four instantiations of non-local blocks. ’input’ denotes the input of the non-local block , ’output’ denotes the output of the non-local block ,’att’ denotes the attention map of query positions .

表1：四种非局部块实例的统计分析。’input’表示非局部块的输入 ，’output’表示非局部块的输出 ，’att’表示查询位置的注意力图 。

# 3.2.2 Statistical Analysis

# 3.2.2 统计分析

To more rigorously verify the phenomenon observed from the visualization, we statistically compare the differences (cosine distances) between the input features and the output features of different positions. Denote as the feature vector for position . The average distance measure is defined as avg\_dist , where is the distance function between two vectors. Cosine distance is a widely-used distance measure, defined as .

为了更严格地验证可视化观察到的现象，我们统计比较了不同位置输入特征和输出特征之间的差异（余弦距离）。记 为位置 的特征向量。平均距离度量定义为 avg\_dist ，其中 是两个向量之间的距离函数。余弦距离是一种广泛使用的距离度量，定义为 。

Different Non-local Instantiations/Tasks. The average cosine distances are computed between input features, attention maps and output features of different positions, with four instantiations of the non-local block on three standard tasks: object detection on COCO, action recognition on Kinetics, and image classification on ImageNet. In detail, we compute the cosine distance between three kinds of vectors, the non-local block inputs ,’input’ in Table 1), the non-local block outputs before fusion , ’output’ in Table 1), and the attention maps of query positions .

不同的非局部实例化/任务。计算了不同位置处的输入特征、注意力图和输出特征之间的平均余弦距离，这四个非局部块的实例化在三个标准任务上：COCO上的对象检测、Kinetics上的动作识别和ImageNet上的图像分类。具体来说，我们计算了三种向量之间的余弦距离，分别是非局部块输入 （表1中的‘input’），融合前非局部块输出 （表1中的‘output’），以及查询位置的注意力图 。

| Dataset | Stage |  |  | cosine distance | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | input | output | att |
| COCO | c3 | 37.5 | 34.4 | 0.326 | 0.004 | 0.009 |
| c4 | 38.1 | 34.8 | 0.401 | 0.012 | 0.020 |
| c5 | 38.2 | 35.1 | 0.372 | 0.024 | 0.042 |
| Kinetics | Stage | Top-1 | Top-5 | input | output | att |
| c3 | 75.5 | 92.1 | 0.297 | 0.007 | 0.005 |
| c4 | 75.4 | 92.1 | 0.395 | 0.001 | 0.001 |
| ImageNet | c3 | 77.0 | 93.4 | 0.248 | 0.067 | 0.041 |
| c4 | 77.2 | 93.5 | 0.301 | 0.074 | 0.115 |
| c5 | 76.5 | 93.2 | 0.257 | 0.013 | 0.033 |

TABLE 2: Statistical analysis of non-local block (Embedded Gaussian) at different stages on four tasks.

表2：非局部块（嵌入式高斯）在不同阶段四个任务上的统计分析。

Results with four instantiations of the non-local block on four standard tasks are shown in Table 1 First, large values of cosine distance in the ’input’ column show that the input features for the non-local block are discriminative across different positions. But the values of cosine distance in the ’output’ column are at least one order of magnitude smaller than that in the ’input’ column on COCO, Kinetics and ImageNet, indicating that output global context features modeled by the non-local block on these three tasks are almost the same for different query positions. The cosine distances on attention maps (’att’) are also very small for all instantiations on these three tasks, which again verifies the observation from the visualization.

表1展示了四个标准任务上非局部块的四种实例化的结果。首先，’input’列中余弦距离的大值表明非局部块的输入特征在不同位置上是可区分的。但是，COCO、Kinetics和ImageNet上’output’列中的余弦距离至少比’input’列中的小一个数量级，表明这些三个任务上由非局部块建模的输出全局上下文特征对于不同的查询位置几乎是相同的。这些三个任务上注意力图（’att’）的余弦距离也非常小，这再次验证了可视化观察到的结果。

To conclude, although a non-local block intends to compute the global context specific to each query position, the global context after training is actually independent of query position. Hence, it may be redundant for the non-local block to compute different attention maps for different positions, allowing us to simplify the non-local block.

总结来说，尽管非局部块的目的是计算每个查询位置特定的全局上下文，但训练后的全局上下文实际上与查询位置无关。因此，非局部块为不同位置计算不同的注意力图可能是多余的，这允许我们简化非局部块。

Insertion at Different Stages. It is widely accepted that the lower layers of deep networks contain low-level, less-semantic features such as local edges, and higher layers contain high-level features with more semantic information, such as parts and objects [62]. The non-local block may perform differently at different places in a deep network. To examine this, we have also done a statistical analysis across different stages with the most widely-used instantiation, Embedded Gaussian, on the four standard tasks.

在不同阶段的插入。普遍认为，深度网络的底层包含较低级别的、较少语义特征，如局部边缘，而更高层则包含具有更多语义信息的高级特征，如部分和对象 [62]。非局部块在深度网络中不同的位置可能表现不同。为了验证这一点，我们还对四个标准任务的最常用实例化方法，嵌入式高斯，在不同阶段进行了统计分析。

For different tasks, the non-local block is applied at different positions. For example, in action recognition on kinetics, the nonlocal blocks are inserted only in c4 and c5 , hence we perform the experiments accordingly.

对于不同的任务，非局部块应用于不同的位置。例如，在 kinetics 上的动作识别中，非局部块只插入在 c4 和 c5 中，因此我们相应地进行实验。

Results are presented in Table 2. Interestingly, we can see an obvious trend from lower layers to higher layers, that the output features in higher layers are more query-dependent than that in the lower layers.

结果展示在表2中。有趣的是，我们可以从底层到高层看到一个明显的趋势，即高层输出的特征比低层的特征更加依赖于查询。

Fine-grained Analysis. To analyze the reason of this phenomenon, we have done a more fine-grained statistical analysis on the two most widely-adopted instantiations of the non-local block, Embedded Gaussian and Gaussian.

精细分析。为了分析这个现象的原因，我们对非局部块的两种最广泛采用的实例化方法，嵌入式高斯和高斯，进行了更精细的统计分析。

For Embedded Gaussian, we compute the average cosine distances between input features (input), features after the transform (key), features after the transform (query), different query features after inner product (prod), attention maps (att), and output features (output), which are marked in Figure 3 (a). For Gaussian, as marked in Figure 3 (b), we compute the average cosine distances between input features (input), different query features after inner product (prod), attention maps (att), and output features (output).

对于嵌入式高斯，我们计算了输入特征（input）、 变换后的特征（key）、 变换后的特征（query）、内积后不同的查询特征（prod）、注意力图（att）和输出特征（output）之间的平均余弦距离，这些在图3（a）中标记。对于高斯，如图3（b）所示，我们计算了输入特征（input）、内积后不同的查询特征（prod）、注意力图（att）和输出特征（output）之间的平均余弦距离。

| Dataset | Method | cosine distance | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | input | key | query | prod | att | output |
| COCO | E-Gaussian | 0.401 | 0.332 | 0.050 | 0.005 | 0.020 | 0.012 |
| Gaussian | 0.397 |  | - | 0.069 | 0.177 | 0.062 |
| Kinetics | E-Gaussian | 0.358 | 0.356 | 0.264 | 0.404 | 0.004 | 0.003 |
| Gaussian | 0.345 | - | - | 0.036 | 0.056 | 0.056 |
| ImageNet | E-Gaussian | 0.301 | 0.234 | 0.156 | 0.340 | 0.115 | 0.074 |
| Gaussian | 0.045 | - | - | 0.001 | 0.011 | 0.005 |

TABLE 3: Fine-grained statistical analysis of non-local block (Embedded Gaussian and Gaussian) on four tasks.

表3：非局部块（嵌入式高斯和高斯）在四个任务上的精细统计分析。

| Method | mloU | input | output | att |
| --- | --- | --- | --- | --- |
| Gaussian | 76.47 | 0.318 | 0.433 | 0.478 |
| E-Gaussian | 77.59 | 0.315 | 0.393 | 0.354 |
| Dot product | 77.74 | 0.323 | 0.386 | 0.331 |
| Concat | 77.78 | 0.321 | 0.002 | 0.001 |

TABLE 4: Statistical analysis using four instantiations of nonlocal blocks on Cityscape semantic segmentation. ’input’ denotes the input of the non-local block ,’output’ denotes the output of the non-local block ,’att’ denotes the attention map of query positions .

表4：使用四种非局部块实例化的城市景观语义分割的统计分析。’input’ 表示非局部块的输入 ，’output’ 表示非局部块的输出 ，’att’ 表示查询位置的注意力图 。

Results of the fine-grained statistical analysis are shown in Table 3 First, we look into the results on COCO, Kinetics and ImageNet. For Embedded Gaussian, although and are both convolutions with the same input, the features after are more similar, and the features after are still different. Also, features after the inner-product computation are more query-independent after training. For Gaussian, as this instantiation does not include the query and key transformations, the attention maps still appear query-dependent. But after attention pooling and the output transform, the differences between the output features are significantly reduced, and are almost one order of magnitude smaller than that of the input features.

细粒度统计分析的结果显示在表3中。首先，我们研究了在COCO、Kinetics和ImageNet上的结果。对于嵌入式高斯，尽管 和 都是具有相同输入的 卷积，但 之后的特征更为相似，而 之后的特征仍然有所不同。此外，内积计算之后的特征在训练后更加查询独立。对于高斯，由于这种实例化不包括查询和键变换，注意力图仍然表现为查询依赖。但是，经过注意力池化和输出变换后，输出特征之间的差异显著减小，几乎比输入特征小一个数量级。

In our understanding, the tasks drive the network components to learn the specific architecture that can benefit the tasks most. And query-independence of the non-local block can benefit three major tasks: object detection on COCO, action recognition on Kinetics, and image recognition on ImageNet.

在我们的理解中，任务驱动网络组件学习能够最有利于任务的特定架构。非局部块的查询独立性可以有利于三大主要任务：在COCO上的目标检测、在Kinetics上的动作识别和在ImageNet上的图像识别。

Exceptions. Although non-local networks do not learn pairwise relations on the above three important visual recognition tasks, we note that there are also some tasks where non-local networks successfully learn pairwise relations, e.g. semantic segmentation on Cityscapes, as illustrated in Table 4 Table 12 also shows that NLNet can improve segmentation accuracy over the regular counterpart. A question is whether such improvements are due mainly to the learnt pairwise relations. Surprisingly, a simplified version of NLNet (noted as SNL, which will be introduced in the next section) which models only global context also shows performance comparable to NLNet. This indicates that although the non-local block applied in semantic segmentation may learn pairwise relations, the accuracy improvement may be mostly ascribed to the modeling of global context.

异常。尽管非局部网络在上述三个重要的视觉识别任务中不学习成对关系，我们注意到也有一些任务中非局部网络成功学习了成对关系，例如在Cityscapes上的语义分割，如表4所示。表12也显示NLNet在常规对比方法上可以提高分割精度。一个问题是这种改进是否主要归因于学习到的成对关系。令人惊讶的是，NLNet的简化版本（记作SNL，将在下一节介绍），它仅模拟全局上下文，也显示出与NLNet相当的性能。这表明尽管在语义分割中应用的非局部块可能学习了成对关系，但精度提升可能主要归因于全局上下文的建模。

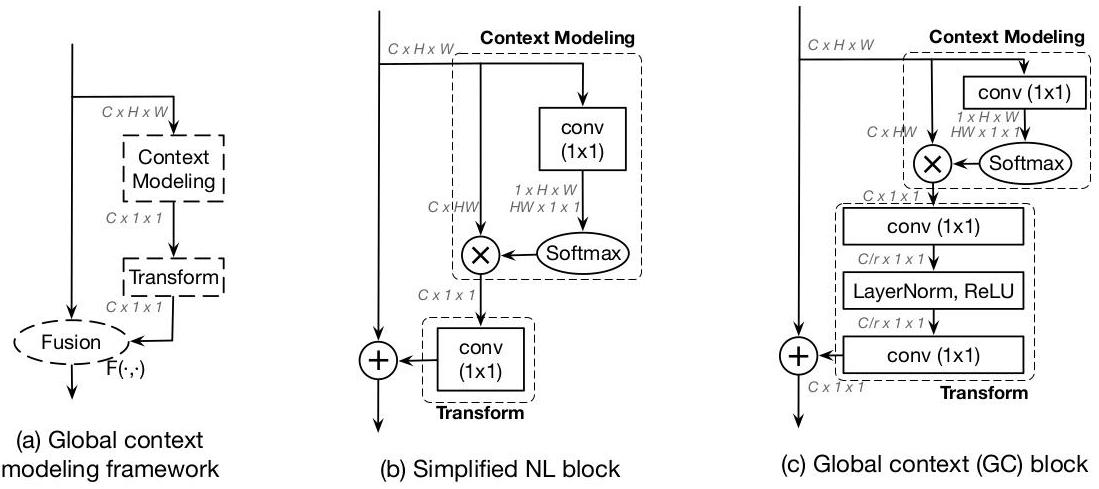


Fig. 4: Architecture of the main blocks. The feature maps are shown as feature dimensions, e.g. CxHxW denotes a feature map with channel number , height and width . denotes matrix multiplication, denotes broadcast element-wise addition, and denotes broadcast element-wise multiplication.

图4：主要模块的结构。特征图显示为特征维度，例如CxHxW表示具有 个通道、 高度和 宽度的特征图。 表示矩阵乘法， 表示广播逐元素加法， 表示广播逐元素乘法。

# 4 Method

# 4 方法

In the last section, both qualitative and statistical analysis indicate that non-local blocks tend to learn query-independent attention maps in many visual recognition tasks, instead of query-dependent context as implied by the formulation. This finding challenges the necessity of the query-dependent formulation in the original non-local block, and raises the question of whether explicit query-independent attention maps perform worse than the original query-dependent formulation. We answer this in the following subsections. We first present a simplified non-local formulation by explicitly making the attention maps query-independent in Section 4.1. We will show in experiments that this simplified formulation can significantly reduce computation yet maintain accuracy. Then in Section 4.2, we abstract this simplified non-local formulation into a general global context modeling framework, which interestingly also operates like the popular SE block [2]. Finally, in Section 4.3, we present our global context block, which is a new instantiation of the general framework by combining the strengths of the simplified non-local block and the SE block [2].

在上一节中，定性和统计分析均表明，在许多视觉识别任务中，非局部块倾向于学习与查询无关的注意力图，而不是查询依赖的上下文，这与原始非局部块的公式所暗示的相反。这一发现质疑了原始非局部块中查询依赖公式必要性的问题，并引发了这样一个问题：显式的查询无关注意力图是否比原始的查询依赖公式表现更差。我们在接下来的小节中回答这个问题。我们首先在4.1节中通过显式地使注意力图与查询无关来提出一个简化的非局部公式。我们将在实验中展示，这种简化的公式可以显著减少计算量，同时保持准确性。然后在4.2节中，我们将这种简化的非局部公式抽象成一个一般的全局上下文建模框架，有趣的是，这个框架也像流行的SE块[2]一样运作。最后，在4.3节中，我们提出了我们的全局上下文块，这是通过结合简化的非局部块和SE块[2]的优点来实现的一般框架的新实例。

# 4.1 Simplifying the Non-local Block

# 4.1 简化非局部块

As the widely-adopted Embedded Gaussian instantiation achieves representative performance on all three standard tasks, as shown in Table 1, we adopt the Embedded Gaussian as the basic nonlocal block in the following sections. Based on the observation that the attention maps for different query positions are almost the same, we simplify the non-local block by computing a global (query-independent) attention map and share this global attention map among all query positions. Following the results in [3] that variants with and without achieve comparable performance, we omit in the simplified version. Our simplified non-local block is defined as

如表1所示，广泛采用的嵌入式高斯实例在所有三个标准任务上均取得了代表性性能，因此我们在以下章节中采用嵌入式高斯作为基本的非局部块。基于不同查询位置的注意力图几乎相同的观察，我们通过计算一个全局（查询无关）的注意力图并让所有查询位置共享这个全局注意力图来简化非局部块。根据[3]中的结果，具有和没有 的变体取得了可比较的性能，我们在简化版本中省略了 。我们的简化非局部块定义为

where and denote linear transformation matrices.

其中 和 表示线性变换矩阵。

To further reduce the computational cost of this simplified block, we apply the distributive law to move outside of the attention pooling, as

为了进一步减少这个简化块的计算成本，我们将分配律应用于将 移出注意力池化，如

This version of simplified non-local block is illustrated in Figure 4(b). After moving outside of attention pooling, the FLOPs of this convolution is reduced from to .

简化的非局部块版本的示意图如图4(b)所示。在将 移出注意力池化后，这种 卷积 的FLOPs从 降低到 。

Different from the traditional non-local block, the second term in Eqn. 3 is independent of the query position , which means that this term is shared across all query positions . We thus directly model global context as a weighted sum of the features at all positions, and aggregate (add) the global context features to the features at each query position. In experiments, we directly replace the non-local (NL) block with our simplified non-local (SNL) block, and evaluate accuracy and computation cost on four tasks, object detection on COCO, semantic segmentation on Cityscapes, ImageNet classification, and action recognition on Kinetics, shown in Tables 5 (a), 8(a), 12(a) and 10. As expected, the SNL block achieves performance comparable to (or slightly below) the NL block with significantly lower FLOPs.

与传统的非局部块不同，方程3中的第二项与查询位置 无关，这意味着这个项在所有查询位置之间是共享的 。因此，我们直接将全局上下文建模为所有位置特征的加权和，并将全局上下文特征聚合（相加）到每个查询位置的特征上。在实验中，我们直接用我们的简化非局部（SNL）块替换非局部（NL）块，并在四个任务上评估准确性和计算成本：COCO上的目标检测、Cityscapes上的语义分割、ImageNet分类以及Kinetics上的动作识别，结果分别显示在表5(a)、8(a)、12(a)和10中。如预期，SNL块在显著降低FLOPs的同时，性能与NL块相当（或略低）。

# 4.2 Global Context Modeling Framework

# 4.2 全局上下文建模框架

As shown in Fig. 4(b), the simplified non-local block can be abstracted into three parts: (a) global attention pooling, which adopts a convolution and a softmax function to obtain the attention weights, and then performs attention pooling to obtain the global context features; (b) feature transform via a convolution ; (c) feature aggregation, which employs addition to aggregate global context features to each position.

如图4(b)所示，简化的非局部块可以抽象为三个部分：(a) 全局注意力池化，它采用 卷积 和softmax函数来获取注意力权重，然后执行注意力池化以获得全局上下文特征；(b) 通过 卷积 进行特征转换；(c) 特征聚合，它使用加法将全局上下文特征聚合到每个位置。

We regard this abstraction as a global context modeling framework, illustrated in Figure 4 (a) and defined as

我们将这种抽象视为全局上下文建模框架，如图4(a)所示，并定义为

where (a) denotes the context modeling module which groups the features of all positions together via weighted averaging with weight to obtain the global context features (global attention pooling in the simplified NL (SNL) block); (b) denotes the feature transform to capture channel-wise dependencies convolution in the SNL block); and (c) denotes the fusion function to aggregate the global context features to the features of each position (broadcast element-wise addition in the SNL block).

其中 (a) 表示上下文建模模块，该模块通过加权平均的方式将所有位置的特征组合在一起，权重为 ，以获得全局上下文特征（简化神经网络（SNL）块中的全局注意力池化）；(b) 表示特征变换，用于捕获通道间的依赖关系（SNL块中的卷积）；(c) 表示融合函数，用于将全局上下文特征聚合到每个位置的特征上（SNL块中的广播逐元素加法）。

Interestingly, the squeeze-excitation (SE) block proposed in [2] is also an instantiation of our proposed framework, which consists of: (a) global average pooling for global context modeling (set in Eqn. 4), called the squeeze operation in the SE block; (b) a bottleneck transform module (let in Eqn. 4 be one convolution, one ReLU, one convolution and a sigmoid function, sequentially), to compute the importance for each channel, called the excitation operation in the SE block; and (c) a rescaling function for fusion (let in Eqn. 4 be element-wise multiplication), to recalibrate the channel-wise features.

有趣的是，文献 [2] 中提出的挤压-激发（SE）块也是我们所提出框架的一个实例，它包括：(a) 用于全局上下文建模的全局平均池化（在公式 4 中设置为 ），在SE块中称为挤压操作；(b) 瓶颈变换模块（令公式 4 中的 为一个 卷积，一个 ReLU，一个 卷积和一个sigmoid函数，依次进行），用于计算每个通道的重要性，在SE块中称为激发操作；(c) 用于融合的重缩放函数（令公式 4 中的 为逐元素乘法），用于重新校准通道特征。

# 4.3 Global Context Block

# 4.3 全局上下文块

Here we propose a new instantiation of the global context modeling framework, named the global context (GC) block, which can effectively model long-range dependency as a simplified non-local block, and is lightweight for application to all layers with a small increase in FLOPs.

在这里，我们提出了全局上下文建模框架的一个新的实例化，命名为全局上下文（GC）块，它能够有效地将长距离依赖建模为一个简化的非局部块，并且轻量级，适用于所有层，FLOPs的增加很小。

In the simplified non-local block, shown in Figure 4(b), the transform module has the largest number of parameters, including from one convolution with parameters. When we add this SNL block to higher layers, e.g. res5, the number of parameters of this convolution, , dominates the number of parameters of this block. Hence, this convolution is replaced by a bottleneck transform module, which significantly reduces the number of parameters from to , where is the bottleneck ratio and denotes the hidden representation dimension of the bottleneck. With the default reduction ratio set to , the number of parameters for the transform module can be reduced to of the original SNL block. More results on different values of bottleneck ratio are shown in Table 5(e).

在简化的非局部块中，如图4(b)所示，变换模块拥有最多的参数，包括一个具有 卷积的 参数。当我们将这个SNL块添加到更高层时，例如res5，这个 卷积的参数数量 ，决定了这个块的总参数数量。因此，这个 卷积被一个瓶颈变换模块所替代，显著地将参数数量从 减少到 ，其中 是瓶颈比例， 表示瓶颈的隐藏表示维度。默认的缩减比例设置为 ，变换模块的参数数量可以减少到原始SNL块的 。

As the two-layer bottleneck transformation increases the difficulty of optimization, we add layer normalization inside the bottleneck transformation (before ReLU) to ease optimization, as well as to act as a regularizer that can benefit generalization. As shown in Table 5(d), layer normalization can significantly enhance the performance of object detection and segmentation on COCO.

由于双层瓶颈变换增加了优化的难度，我们在瓶颈变换内部（ReLU之前）添加了层归一化以简化优化，同时作为正则化器，有助于泛化。如表5(d)所示，层归一化可以显著提高在COCO上进行目标检测和分割的性能。

The detailed architecture of the global context (GC) block is illustrated in Figure 4(c) and formulated as

全局上下文（GC）块的具体架构如图4(c)所示，并表述为

where is the weight for global attention pooling, and denotes the bottleneck transform. Specifically, our GC block consists of: (a) global attention pooling for context modeling; (b) bottleneck transform to capture channel-wise dependencies; and (c) broadcast element-wise addition for feature fusion.

其中 是全局注意力池化的权重， 表示瓶颈变换。具体来说，我们的GC块包括：(a) 用于上下文建模的全局注意力池化；(b) 用于捕获通道依赖的瓶颈变换；(c) 用于特征融合的广播逐元素加法。

Since the GC block is lightweight, it can be applied in multiple layers to better capture long-range dependency with only a slight increase in computation cost. Taking ResNet-50 for ImageNet classification as an example, GC-ResNet-50 denotes adding the GC block to all layers (c3+c4+c5) in ResNet-50 with a bottleneck ratio of 16. GC-ResNet-50 increases ResNet-50 computation from GFLOPs to GFLOPs, corresponding to a relative increase. Also, GC-ResNet-50 introduces additional parameters beyond the parameters required by ResNet- 50, corresponding to a increase.

由于 GC 块轻量级，它可以被应用于多层以更好地捕获长距离依赖，同时计算成本仅略有增加。以 ImageNet 分类中的 ResNet-50 为例，GC-ResNet-50 表示将 GC 块添加到 ResNet-50 的所有层（c3+c4+c5）中，瓶颈比为 16。GC-ResNet-50 将 ResNet-50 的计算量从 GFLOPs 增加到 GFLOPs，对应 的相对增加。此外，GC-ResNet-50 引入了 的额外参数，超过了 ResNet-50 所需的 参数，对应 的增加。

Global context can benefit a wide range of visual recognition tasks, and the flexibility of the GC block allows it to be plugged into network architectures used in various computer vision problems. In this paper, we apply our GC block to four general vision tasks - image recognition, object detection/instance segmentation, semantic segmentation and action recognition - and observe significant improvements in all four.

全局上下文可以受益于广泛的视觉识别任务，GC 块的灵活性使其可以被插入到用于各种计算机视觉问题的网络架构中。在本文中，我们将 GC 块应用于四个通用视觉任务 - 图像识别、对象检测/实例分割、语义分割和动作识别 - 并在所有四个任务中观察到显著的改进。

Relationship to non-local block. As the non-local block actually learns query-independent global context, the global attention pooling of our global context block models the same global context as the NL block but with significantly lower computation cost. As the GC block adopts the bottleneck transform to reduce redundancy in the global context features, the number of parameters and FLOPs are further reduced. The FLOPs and number of parameters of the GC block are significantly lower than that of the NL block, allowing our GC block to be applied to multiple layers with just a slight increase in computation, while better capturing long-range dependency and aiding network training.

与非局部块的关系。由于非局部块实际上学习查询无关的全局上下文，我们全局上下文块的全局注意力池化与非局部块模型相同的全局上下文，但计算成本显著降低。由于 GC 块采用瓶颈变换来减少全局上下文特征中的冗余，参数数量和 FLOPs 进一步减少。GC 块的 FLOPs 和参数数量显著低于 NL 块，使得我们的 GC 块可以被应用于多层，而计算量仅略有增加，同时更好地捕获长距离依赖并帮助网络训练。

Relationship to squeeze-excitation block. The main difference between the SE block and our GC block is the fusion module, which reflects the different goals of the two blocks. The SE block adopts rescaling to recalibrate the importance of channels but inadequately models long-range dependency. Our GC block follows the NL block by utilizing addition to aggregate global context to all positions for capturing long-range dependency. A second difference is with the layer normalization in the bottleneck transform. As our GC block adopts addition for fusion, layer normalization can ease optimization of the two-layer architecture for the bottleneck transform, which can lead to better performance. Third, global average pooling in the SE block is a special case of global attention pooling in the GC block. Results in Tables 5(d), 5(f) and 8(b) show the superiority of addition in the fusion module, layer normalization in the two-layer bottleneck, and the global attention pooling, compared to the SE block, respectively.

与挤压-激励块的关系。SE块与我们的GC块之间的主要区别在于融合模块，这反映了两个块的不同目标。SE块采用重新缩放来重新校准通道的重要性，但不足以建模长距离依赖。我们的GC块通过使用加法来汇总全局上下文到所有位置，以捕获长距离依赖，遵循NL块。第二个区别在于瓶颈变换中的层归一化。由于我们的GC块采用加法进行融合，层归一化可以简化双层架构的优化，从而可能导致更好的性能。第三，SE块中的全局平均池化是GC块中全局注意力池化的一个特例。表5(d)、5(f)和8(b)的结果分别显示了融合模块中的加法、双层瓶颈中的层归一化和全局注意力池化相对于SE块的优势。

# 5 EXPERIMENTS

# 5 实验部分

To evaluate the proposed method, we carry out experiments on four basic tasks: object detection/instance segmentation on COCO [63], image classification on ImageNet [64], action recognition on Kinetics [65], and semantic segmentation on Cityscapes [66]. Experimental results demonstrate that the proposed GCNet generally outperforms non-local networks with significantly lower FLOPs.

为了评估所提出的方法，我们在四个基本任务上进行了实验：COCO [63] 上的目标检测/实例分割、ImageNet [64] 上的图像分类、Kinetics [65] 上的动作识别以及Cityscapes [66] 上的语义分割。实验结果表明，提出的GCNet通常在显著降低的FLOPs下优于非局部网络。

# 5.1 Object Detection/Instance Segmentation on COCO

# 5.1 COCO上的目标检测/实例分割

We investigate our model on object detection and instance segmentation on COCO 2017 [63], whose train set is comprised of images, validation set of images, and test-dev set of images. We follow the standard setting [7] of evaluating object detection and instance segmentation via the standard mean average-precision scores at different boxes and the mask IoUs.

我们在COCO 2017 [63]上的目标检测和实例分割任务上研究我们的模型，其训练集包含 张图像，验证集包含 张图像，测试开发集包含 张图像。我们遵循标准设置 [7]，通过不同框的标准平均精度分数和掩膜IoU来评估目标检测和实例分割。

Setup. Our experiments are implemented with PyTorch [67] based on open source mmdetection [68]. Unless otherwise noted, our GC block of ratio is applied to stages c3, c4, c5 of ResNet/ResNeXt.

设置。我们的实验使用基于开源mmdetection [68]的PyTorch [67]进行实现。除非另有说明，我们的GC块的比例 应用于ResNet/ResNeXt的c3、c4、c5阶段。

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

IEEE Transactions on Pattern Analysis and Machine Intelligence

|  | (a) Block design | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Apmask |  | A Dmask | #param | FLOPs |
| baseline | 37.2 | 59.0 | 40.1 | 33.8 | 55.4 | 35.9 | 44.4M | 279.4G |
| +1 NL | 38.0 | 59.8 | 41.0 | 34.7 | 56.7 | 36.6 | 46.5M | 288.7G |
| +1 SNL | 38.1 | 60.0 | 41.6 | 35.0 | 56.9 | 37.0 | 45.4M | 279.4G |
| +1 GC | 38.1 | 60.0 | 41.2 | 34.9 | 56.5 | 37.2 | 44.5M | 279.4G |
| +all GC | 39.4 | 61.6 | 42.4 | 35.7 | 58.4 | 37.6 | 46.9M | 279.6G |

|  | (d) Bottleneck design | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | A pmask | #param | FLOPs |
| baseline | 37.2 | 59.0 | 40.1 | 33.8 | 55.4 | 35.9 | 44.4M | 279.4G |
| w/o ratio | 39.4 | 61.8 | 42.8 | 35.9 | 58.6 | 38.1 | 64.4M | 279.6G |
| r16 (ratio 16) | 38.8 | 61.0 | 42.3 | 35.3 | 57.6 | 37.5 | 46.9M | 279.6G |
| r16+ReLU | 38.8 | 61.0 | 42.0 | 35.4 | 57.5 | 37.6 | 46.9M | 279.6G |
| r16+LN+ReLU | 39.4 | 61.6 | 42.4 | 35.7 | 58.4 | 37.6 | 46.9M | 279.6G |

(e) Bottleneck ratio

(e) 瓶颈比例

|  |  |  |  | APmask |  |  | #param | FLOPs |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| baseline | 37.2 | 59.0 | 40.1 | 33.8 | 55.4 | 35.9 | 44.4M | 279.4G |
| ratio 4 | 39.9 | 62.2 | 42.9 | 36.2 | 58.7 | 38.3 | 54.4M | 279.6G |
| ratio 8 | 39.5 | 62.1 | 42.5 | 35.9 | 58.1 | 38.1 | 49.4M | 279.6G |
| ratio 16 | 39.4 | 61.6 | 42.4 | 35.7 | 58.4 | 37.6 | 46.9M | 279.6G |
| ratio 32 | 39.1 | 61.6 | 42.4 | 35.7 | 58.1 | 37.8 | 45.7M | 279.5G |

|  |  |  |  | APmask |  |  | #param | FLOPs |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| baseline | 37.2 | 59.0 | 40.1 | 33.8 | 55.4 | 35.9 | 44.4M | 279.4G |
| avg+scale | 38.2 | 60.2 | 41.2 | 34.7 | 56.7 | 37.1 | 46.9M | 279.5G |
| avg+add | 39.1 | 61.4 | 42.3 | 35.6 | 57.9 | 37.9 | 46.9M | 279.5G |
| att+scale | 38.3 | 60.4 | 41.5 | 34.8 | 57.0 | 36.8 | 46.9M | 279.6G |
| att+add | 39.4 | 61.6 | 42.4 | 35.7 | 58.4 | 37.6 | 46.9M | 279.6G |

TABLE 5: Ablation study based on Mask R-CNN, using ResNet- 50 as backbone with FPN, for object detection and instance segmentation on COCO 2017 validation set.

表5：基于Mask R-CNN的消融研究，使用ResNet-50作为主干网络并带有FPN，针对COCO 2017验证集进行目标检测和实例分割。

Training. We use the standard configuration of Mask R-CNN [7] with FPN and ResNet/ResNeXt as the backbone architecture. The input images are resized such that their shorter side is of 800 pixels [69]. We trained on 8 GPUs with 2 images per GPU (effective mini batch size of 16). The backbones of all models are pretrained on ImageNet classification [64], then all layers except for and are jointly finetuned with detection and segmentation heads. Unlike stage-wise training with respect to RPN in [7], end-to-end training like in [70] is adopted for our implementation, yielding better results. Different from the conventional finetuning setting [7], we use Synchronized BatchNorm to replace frozen BatchNorm. All models are trained for 12 epochs using Synchronized SGD with a weight decay of 0.0001 and momentum of 0.9 , which roughly corresponds to the schedule in the Mask R-CNN benchmark [71]. The learning rate is initialized to 0.02 , and decays by a factor of 10 at the 8th and 11th epochs. The choice of hyper-parameters also follows the latest release of the Mask R-CNN benchmark [71].

训练。我们使用带有FPN和ResNet/ResNeXt作为主干架构的Mask R-CNN [7]的标准配置。输入图像的大小调整为其较短的边为800像素 [69]。我们在8个GPU上训练，每个GPU处理2张图像（有效的最小批量大小为16）。所有模型的骨干网络在ImageNet分类 [64]上进行预训练，然后除了 和 之外的所有层都与检测和分割头共同微调。与[7]中关于RPN的逐阶段训练不同，我们的实现采用了类似于[70]的端到端训练，从而获得更好的结果。与传统微调设置 [7] 不同，我们使用同步批量归一化来替换冻结的批量归一化。所有模型都使用同步SGD训练12个周期，权重衰减为0.0001，动量为0.9，这大致对应于Mask R-CNN基准 [71]中的 计划。初始学习率设置为0.02，并在第8和第11个周期时以10倍因子衰减。超参数的选择也遵循Mask R-CNN基准 [71]的最新版本。

# 5.1.1 Ablation Study

# 5.1.1 消融研究

The ablation study is done on the COCO 2017 validation set. The standard COCO metrics including for both bounding boxes and segmentation masks are reported.

消融研究在 COCO 2017 验证集上进行。报告了标准 COCO 指标，包括 用于边界框和分割掩膜。

| (a) Different Normalization | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| backbone | head | method | APbbox |  |  | APmask |  |  |
| fixBN | (w/o BN) | baseline | 37.3 | 59.0 | 40.2 | 34.2 | 55.9 | 36.2 |
| +GC r16 | 38.5 | 60.8 | 41.5 | 35.1 | 57.3 | 37.1 |
| +GC r4 | 38.9 | 61.1 | 42.0 | 35.5 | 57.7 | 37.5 |
| syncBN | 2fc (w/o BN) | baseline | 37.2 | 59.0 | 40.1 | 33.8 | 55.4 | 35.9 |
| +GC r16 | 39.4 | 61.6 | 42.4 | 35.7 | 58.4 | 37.6 |
| +GC r4 | 39.9 | 62.2 | 42.9 | 36.2 | 58.7 | 38.3 |
| syncBN | 4conv1fc syncBN | baseline | 38.8 | 59.5 | 42.6 | 34.6 | 56.2 | 37.1 |
| +GC r16 | 41.0 | 62.1 | 44.9 | 36.5 | 58.3 | 39.0 |
| +GC r4 | 41.4 | 62.5 | 45.5 | 37.0 | 59.1 | 39.5 |
| setting | (b) Longer Training | | | | | | | |
| schd | method | APbbox |  |  | APmask |  |  |
| syncBN 2fc | 2x | baseline | 37.7 | 59.1 | 40.9 | 34.3 | 55.8 | 36.5 |
| +GC r16 | 39.7 | 61.8 | 43.0 | 36.0 | 58.5 | 38.4 |
| +GC r4 | 40.2 | 62.2 | 43.5 | 36.3 | 58.6 | 38.5 |

TABLE 6: Ablation study on different normalization and training schedules for object detection and instance segmentation on COCO 2017 validation set.

表 6：在 COCO 2017 验证集上对对象检测和实例分割的不同归一化和训练计划进行的消融研究。

Block design. Following [5], we insert 1 non-local block (NL), 1 simplified non-local block (SNL), or 1 global context block (GC) right before the last residual block of c4. Table 5 (a) shows that both SNL and GC achieve performance comparable to NL with fewer parameters and less computation, indicating redundancy in computation and parameters in the original non-local design. Furthermore, adding the GC block in all residual blocks yields higher performance on and with a slight increase in FLOPs and #params.

块设计。遵循 [5]，我们在 c4 的最后一个残差块之前插入 1 个非局部块 (NL)、1 个简化非局部块 (SNL) 或 1 个全局上下文块 (GC)。表 5 (a) 显示，SNL 和 GC 以更少的参数和更少的计算量实现了与 NL 相当的性能，表明原始非局部设计中计算和参数存在冗余。此外，在所有残差块中添加 GC 块在 上 和 取得了更高的性能，FLOPs 和参数数量仅略有增加。

Positions. The NL block is inserted after the residual block (afterAdd), while the SE block is integrated after the last convolution inside the residual block (after1x1). In Table 5 (b), we investigate both cases with the GC block and they yield similar results. Hence, we adopt after 1x1 as the default.

位置。NL 块在残差块之后（afterAdd）插入，而 SE 块在残差块内的最后一个 卷积之后（after1x1）集成。在表 5 (b) 中，我们研究了 GC 块在这两种情况下的结果，它们产生了相似的结果。因此，我们采用 after 1x1 作为默认设置。

Stages. Table 5(c) shows the results of integrating the GC block at different stages. All stages benefit from global context modeling in the GC block on and ). Inserting into c4 and c5 both achieves better performance than into c3, demonstrating that better semantic features can benefit more from the global context modeling. With a slight increase in FLOPs, inserting the GC block into all layers (c3+c4+c5) yields even higher performance than inserting into only a single layer.

阶段。表 5(c) 显示了在不同阶段集成 GC 块的结果。所有阶段都从 GC 块的全局上下文建模中受益 在 和 上）。在 c4 和 c5 中插入均比在 c3 中插入取得了更好的性能，表明更好的语义特征能从全局上下文建模中获益更多。在 FLOPs 略有增加的情况下，将 GC 块插入所有层（c3+c4+c5）比仅插入单一层取得了更高的性能。

Bottleneck design. The effects of each component in the bottleneck transform are shown in Table 5(d). w/o ratio denotes the simplified NLNet using one convolution as the transform, which has more parameters compared to the baseline. Even though r16 and r16+ReLU have much fewer parameters than the w/o ratio variant, two layers are found to be harder to optimize and lead to worse performance than a single layer. So LayerNorm (LN) is exploited to ease optimization, leading to performance similar to w/o ratio but with much fewer #params.

瓶颈设计。瓶颈变换中每个组件的效果在表5(d)中显示。w/o ratio表示使用一个 卷积作为变换的简化的NLNet，其参数数量比基线多。尽管r16和r16+ReLU的参数数量比w/o ratio变体少得多，但发现两层更难以优化，并且性能比单层差。因此，利用LayerNorm（LN）来简化优化，使得性能与w/o ratio相似，但参数数量大大减少。

The reason we adopt layer norm here is that other alternatives, i.e. batch norm and group norm, do not perform well probably due to insufficient statistics to compute the means and variances. The spatial resolution of the intermediate feature map in the GC block has been reduced to (see Fig 4(c)). If batch normalization is used, the number of elements to compute each mean and variance is ( is the batch size), which is small. If group normalization is used, the number of elements to compute each mean and variance is ( is the group number), which is also small. For layer norm, the number of elements used to compute each mean and variance is , which is observed to be sufficient.

我们在这里采用层归一化的原因是其他替代方法，即批量归一化和分组归一化，由于计算均值和方差时统计信息不足，可能表现不佳。GC块中的中间特征图的空间分辨率已降低到 （见图4(c)）。如果使用批量归一化，计算每个均值和方差使用的元素数量是 （ 是批量大小），这个数字很小。如果使用分组归一化，计算每个均值和方差使用的元素数量是 （ 是组数），这个数字也很小。对于层归一化，计算每个均值和方差使用的元素数量是 ，观察到这是足够的。

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

IEEE Transactions on Pattern Analysis and Machine Intelligence

| (a) test on validation set | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| backbone | |  | | | FLOPS | | | |
| R50 | baseline | 37.2 | 59.0 | 40.1 | 33.8 | 55.4 | 35.9 | 279.4G |
| +GC r16 | 39.4 | 61.6 | 42.4 | 35.7 | 58.4 | 37.6 | 279.6G |
| +GC r4 | 39.9 | 62.2 | 42.9 | 36.2 | 58.7 | 38.3 | 279.6G |
| R101 | baseline | 39.8 | 61.3 | 42.9 | 36.0 | 57.9 | 38.3 | 354.0G |
| +GC r16 | 41.1 | 63.6 | 45.0 | 37.4 | 60.1 | 39.6 | 354.3G |
| +GC r4 | 41.7 | 63.7 | 45.5 | 37.6 | 60.5 | 39.8 | 354.3G |
| X101 | baseline | 41.2 | 63.0 | 45.1 | 37.3 | 59.7 | 39.9 | 357.9G |
| +GC r16 | 42.4 | 64.6 | 46.5 | 38.0 | 60.9 | 40.5 | 358.2G |
| +GC r4 | 42.9 | 65.2 | 47.0 | 38.5 | 61.8 | 40.9 | 358.2G |
| X101 +Cascade | baseline | 44.7 | 63.0 | 48.5 | 38.3 | 59.9 | 41.3 | 536.9G |
| +GC r16 | 45.9 | 64.8 | 50.0 | 39.3 | 61.8 | 42.1 | 537.2G |
| +GC r4 | 46.5 | 65.4 | 50.7 | 39.7 | 62.5 | 42.7 | 537.3G |
| X101+DCN +Cascade | baseline | 47.1 | 66.1 | 51.3 | 40.4 | 63.1 | 43.7 | 547.5G |
| +GC r16 | 47.9 | 66.9 | 52.2 | 40.9 | 63.7 | 44.1 | 547.8G |
| +GC r4 | 47.9 | 66.9 | 51.9 | 40.8 | 64.0 | 44.0 | 547.8G |
| X101 64x4d+DCN+Cascade + 4conv1fc head + multiscale + | +GC r4 3x schd | 51.8 | 70.4 | 56.1 | 44.7 | 67.9 | 48.4 | 1040.6G |
| (b) test on test-dev set | | | | | | | | |
| X101 +Cascade | baseline | 45.0 | 63.7 | 49.1 | 38.7 | 60.8 | 41.8 | 536.9G |
| +GC r16 | 46.5 | 65.7 | 50.7 | 40.0 | 62.9 | 43.1 | 537.2G |
| +GC r4 | 46.6 | 65.9 | 50.7 | 40.1 | 62.9 | 43.3 | 537.3G |
| X101+DCN +Cascade | baseline | 47.7 | 66.7 | 52.0 | 41.0 | 63.9 | 44.3 | 547.5G |
| +GC r16 | 48.3 | 67.5 | 52.7 | 41.5 | 64.6 | 45.0 | 547.8G |
| +GC r4 | 48.4 | 67.6 | 52.7 | 41.5 | 64.6 | 45.0 | 547.8G |
| X101 64x4d+DCN+Cascade + 4conv1fc head + multiscale + | +GC r4 3x schd | 52.3 | 70.9 | 56.9 | 45.4 | 68.9 | 49.6 | 1040.6G |

TABLE 7: Results of GCNet (ratio 4 and 16) with stronger backbones on COCO 2017 validation and test-dev sets.

表7：在COCO 2017验证集和测试-dev集上，具有更强基础网络的GCNet（比例4和16）的结果。

Bottleneck ratio. The bottleneck design is intended to reduce redundancy in parameters and provide a good tradeoff between performance and #params. In Table 5(e), we vary the ratio of the bottleneck. As the ratio decreases (from 32 to 4) with increasing number of parameters and FLOPs, the performance improves consistently on and on ), indicating that our bottleneck strikes a good balance between performance and number of parameters. It is worth noting that even with a ratio of , the network still outperforms baseline by large margins.

瓶颈比例。瓶颈设计旨在减少参数冗余，并在性能和参数数量之间提供良好的权衡。在表5(e)中，我们改变了瓶颈的比例 。随着比例 的减小（从32到4），参数数量和FLOPs的增加，性能在 上持续提高 和 上 )，表明我们的瓶颈在性能和参数数量之间取得了良好的平衡。值得注意的是，即使比例达到 ，网络仍然比基线有大幅度的性能提升。

Pooling and fusion. The different choices for pooling and fusion are ablated in Table 5 f). First, it shows that addition is more effective than scaling in the fusion stage. It is surprising that attention pooling only achieves slightly better results than vanilla average pooling. This indicates that how global context is aggregated to query positions (choice of fusion module) is more important than how features from all positions are grouped together (choice in context modeling module). It is worth noting that att+add significantly outperforms avg+scale, which denotes the approach of SENet with layer norm, because of the effective modeling of long-range dependency with attention pooling for context modeling, and the use of addition for feature aggregation.

池化和融合。表5(f)展示了池化和融合不同选择的消融实验。首先，它表明在融合阶段，加法比缩放更有效。令人惊讶的是，注意力池化仅比普通的平均池化略微提高了一些结果。这表明全局上下文如何聚集到查询位置（融合模块的选择）比所有位置的特征如何组合在一起（上下文建模模块的选择）更重要。值得注意的是，att+add 显著优于 avg+scale，后者表示带有层归一化的SENet方法，因为注意力池化在上下文建模中有效建模了长距离依赖，并且使用了加法进行特征聚合。

Different Normalization The result of different normalization is presented in 6(a). GCNet improves the performance by on and on by replacing fixBN with syncBN in the backbone, while baseline maintains similar performance. Since the backbone is already pretrained on ImageNet while the inserted GC block is randomly initialized, the running statistics of the backbone features could help with the training of the GC block. Following [72], [73], syncBN is further applied in both the backbone and heads. Even though the baseline improves by in and in , the gap between GC and the baseline is still preserved, which is in and

不同的归一化方法的结果展示在6(a)中。GCNet通过在主干网络中用syncBN替换fixBN，在 上提高了 的性能，在 上提高了 的性能，而基线模型保持了相似的性能。由于主干网络已经在ImageNet上预训练，而插入的GC模块是随机初始化的，因此主干网络的特征运行统计可能有助于GC模块的训练。遵循[72]，[73]，在主干网络和头部进一步应用syncBN。尽管基线在 上提高了 ，在 上提高了 ，但GC与基线之间的差距仍然保持，分别是 在 上。

(a) Block Design

(a) 块设计

|  | Top-1 Acc | Top-5 Acc | #params(M) |  |
| --- | --- | --- | --- | --- |
| baseline | 76.51 | 93.35 | 25.56 | 3.86 |
| +1NL | 77.21 | 93.64 | 27.66 | 4.11 |
| +1SNL | 77.10 | 93.56 | 26.61 | 3.86 |
| +1GC | 77.20 | 93.47 | 25.69 | 3.86 |
| +all GC | 77.49 | 93.67 | 28.08 | 3.87 |

(b) Pooling and fusion

(b) 池化和融合

|  | Top-1 Acc | Top-5 Acc | #params(M) | FLOPs(G) |
| --- | --- | --- | --- | --- |
| baseline | 76.51 | 93.35 | 25.56 | 3.86 |
| avg+scale | 77.14 | 93.57 | 28.07 | 3.87 |
| avg+add | 77.16 | 93.63 | 28.07 | 3.87 |
| att+scale | 77.18 | 93.58 | 28.08 | 3.87 |
| att+add | 77.49 | 93.67 | 28.08 | 3.87 |

TABLE 8: Ablation study of GCNet with ResNet-50 for image classification on ImageNet validation set.

表8：使用ResNet-50在ImageNet验证集上进行图像分类的GCNet消融研究。

# in .

# 在 中的变化。

Longer Training We also trained our model for 24 epochs which is roughly the same as the schedule in [71]. As shown in 6(b), GCNet does not saturate and greater performance gain is observed, which indicates the large potential capacity of GCNet.

更长的训练时间 我们还将我们的模型训练了24个周期，这大致与[71]中的 计划相同。如6(b)所示，GCNet没有饱和，并且观察到了更大的性能增益，这表明GCNet具有巨大的潜力容量。

# 5.1.2 Experiments on Stronger Backbones

# 5.1.2 在更强主干网络上的实验

We evaluate our GCNet on stronger backbones, by replacing ResNet-50 with ResNet-101 and ResNeXt-101 [15], adding deformable convolution to multiple layers and adopting the Cascade strategy [74]. The results of our GCNet with GC blocks integrated in all layers with bottleneck ratios of 4 and 16 are reported. Table 7 (a) presents detailed results on the validation set. It is worth noting that even when adopting stronger backbones, the gain of GCNet compared to the baseline is still significant, which demonstrates that our GC block with global context modeling is complementary to the capacity of current models. For the strongest backbone, with deformable convolution and cascade RCNN in ResNeXt-101, our GC block can still boost performance by on and on . To further evaluate our proposed method, the results on the test-dev set are also reported, shown in Table 7(b). On test-dev, strong baselines are also boosted by large margins by adding GC blocks, which is consistent with the results on the validation set. These results demonstrate the robustness of our proposed method.

我们在更强的主干网络上评估了我们的GCNet，通过将ResNet-50替换为ResNet-101和ResNeXt-101 [15]，在多层添加可变形卷积 并采用级联策略 [74]。在所有层集成GC模块 且瓶颈比为4和16的GCNet结果已报告。表7(a)展示了验证集上的详细结果。值得注意的是，即使采用更强的主干网络，GCNet相较于基线的提升仍然是显著的，这表明我们的GC模块与全局上下文建模对当前模型的容量具有互补作用。对于最强的主干网络，在ResNeXt-101中采用可变形卷积和级联RCNN，我们的GC模块仍然能够提高 在 上的性能和 在 上的性能。为了进一步评估我们提出的方法，表7(b)还报告了测试集(test-dev)上的结果。在test-dev上，通过添加GC模块，强基线也得到了大幅提升，这与验证集上的结果一致。这些结果证明了我们提出方法的鲁棒性。

# 5.2 Image Classification on ImageNet

# 5.2 在ImageNet上的图像分类

ImageNet [64] is a benchmark dataset for image classification, containing training images and validation images from 1000 classes. We follow the standard setting in [8] to train deep networks on the training set and report the single-crop top- 1 and the top-5 errors on the validation set. Our preprocessing and augmentation strategy follows the baseline proposed in [75] and [2]. Concretely, the following augmentation and preprocessing are performed sequentially during training: random rotation, random aspect ratio with random area crop, resizing, horizontally flip with 0.5 probability, HSV random scaling, and PCA noise sampled from . The standard ResNet-50 is trained for 120 epochs on 4 GPUs with 64 images per GPU (effective batch size of 256) with synchronous SGD of momentum 0.9 . Cosine learning rate decay is adopted with an initial learning rate of 0.1 .

ImageNet [64] 是图像分类的基准数据集，包含 训练图像和 验证图像，分为 1000 个类别。我们遵循 [8] 中的标准设置，在训练集上训练深度网络，并在验证集上报告单裁剪 top-1 和 top-5 错误。我们的预处理和增强策略遵循 [75] 和 [2] 提出的基线。具体来说，以下增强和预处理操作在训练过程中依次执行： 随机旋转， 随机宽高比以及 随机区域裁剪， 调整大小，以 0.5 的概率水平翻转， HSV 随机缩放，以及从 采样的 PCA 噪声。

Block Design. As done for the block design on COCO, results on different blocks are reported in Table 8 (a). The GC block performs slightly better than the NL and SNL blocks with fewer parameters and less computation, which indicates the versatility and generalization ability of our design. By inserting GC blocks in all residual blocks , the performance is further boosted (by on top-1 accuracy compared to baseline) with marginal computational overhead relative increase on FLOPs). In comparison to the baseline, a GC block requires about less computation than an NL block, i.e. vs. , which is significant.

块设计。与 COCO 上的块设计一样，不同块的结果在表 8（a）中报告。GC 块在参数更少、计算量更小的情况下，性能略优于 NL 和 SNL 块，这表明我们的设计具有通用性和泛化能力。通过在所有残差块中插入 GC 块 ，性能进一步提升（相对于基线，top-1 准确度提高 ，计算开销边际增加 （相对增加的 FLOPs））。与基线相比，GC 块需要的计算量大约比 NL 块少 ，即 与 之间的差异是显著的。

|  | Top-1 Acc | Top-5 Acc | #params(M) |  |
| --- | --- | --- | --- | --- |
| baseline | 76.51 | 93.35 | 25.56 | 3.86 |
| SENet\* [2] | 76.86 | 93.30 | 28.07 | 3.87 |
| CBAM\* [47] | 77.34 | 93.69 | 28.07 | 3.87 |
| GCNet | 77.49 | 93.67 | 28.08 | 3.87 |

TABLE 9: Comparison of state-of-the-art methods with ResNet- 50 for image classification on ImageNet validation set. \* denotes that the results are directly taken from the original paper.

表9：将最先进的方法与ResNet-50在ImageNet验证集上进行图像分类的比较。\*表示结果直接取自原论文。

| method | Top-1 Acc | Top-5 Acc | #params(M) | FLOPs(G) |
| --- | --- | --- | --- | --- |
| baseline | 74.94 | 91.90 | 32.45 | 39.29 |
| +5 NL | 75.95 | 92.29 | 39.81 | 59.60 |
| +5 SNL | 75.76 | 92.44 | 36.13 | 39.32 |
| +5 GC | 75.85 | 92.25 | 34.30 | 39.31 |
| +all GC | 76.00 | 92.34 | 42.45 | 39.35 |

TABLE 10: Results of GCNet and NLNet based on Slow-only baseline using R50 as backbone on Kinetics validation set.

表10：基于Slow-only基线使用R50作为主干网络在Kinetics验证集上GCNet和NLNet的结果。

Pooling and fusion. The functionality of different pooling and fusion methods is also investigated on image classification. Comparing Table 8 b) with Table 5 (f), it is seen that attention pooling has greater effect in image classification, which could be one of the missing ingredients in [2]. Also, attention pooling with addition (GCNet) outperforms vanilla average pooling with scaling (SENet with layer norm) by on top-1 accuracy with almost the same #params and FLOPs.

池化和融合。本文还研究了不同池化和融合方法在图像分类上的功能。比较表8 b)与表5 (f)，可以看出注意力池化在图像分类中具有更大的效果，这可能是[2]中缺失的要素之一。此外，带有加法的注意力池化（GCNet）在top-1准确度上超过了带有缩放的普通平均池化（带有层归一化的SENet） ，参数数量和FLOPs几乎相同。

Comparison with Other Approaches. As shown in Table 9, we compare our approach with other state-of-the-art approaches on image recognition of ImageNet, and find that our GCNet outperforms SENet [2] and CBAM [47].

与其他方法的比较。如表9所示，我们将我们的方法与其他最先进的方法在ImageNet的图像识别上进行比较，发现我们的GCNet优于SENet [2]和CBAM [47]。

# 5.3 Action Recognition on Kinetics

# 5.3 动作识别在Kinetics数据集上

For human action recognition, we adopt the widely-used Kinetics [65] dataset, which has training videos and validation videos in 400 human action categories. All models are trained on the training set and tested on the validation set. Following [5], we report top-1 and top-5 recognition accuracy. We adopt the slow-only baseline in [9], the best single model to date that can utilize weights inflated [39] from the ImageNet pretrained model. The inflated 3D strategy [5] greatly speeds up convergence compared to training from scratch. All the experiment settings follow [9]; the slow-only baseline is trained with 8 frames as input, and multi(30)-clip validation is adopted.

对于人体动作识别，我们采用了广泛使用的Kinetics [65] 数据集，该数据集包含 个训练视频和 个验证视频，分为400个人体动作类别。所有模型都在训练集上训练，在验证集上测试。遵循[5]，我们报告了top-1和top-5识别准确度。我们采用了[9]中的slow-only基线，这是迄今为止能够利用ImageNet预训练模型膨胀[39]权重的最佳单一模型。膨胀的3D策略[5]与从头开始训练相比，大大加快了收敛速度。所有实验设置都遵循[9]；slow-only基线以8帧 作为输入，并采用multi(30)-clip验证。

| method | Top-1 Acc | Top-5 Acc | #params(M) | FLOPs(G) |
| --- | --- | --- | --- | --- |
| Slow-only [9] | 74.94 | 91.90 | 32.45 | 39.29 |
| GloRE\* [49] | 75.12 | - |  | 28.90 |
| NLNet [5] | 75.95 | 92.29 | 39.81 | 59.60 |
| GCNet | 76.00 | 92.34 | 42.45 | 39.35 |

TABLE 11: Comparison of state-of-the-art methods with R50 as backbone on Kinetics validation set. \* denotes that the results are directly taken from the original paper. The GloRE results are based on lite version of C2D, and thus have lower FLOPs.

表11：在Kinetics验证集上，使用R50作为主干网络的最新方法比较。\*表示结果直接取自原论文。GloRE结果基于C2D的轻量级版本，因此具有较低的FLOPs。

Ablation Study. The ablation study results are reported in Table 10 For the Kinetics experiments, the ratio of GC blocks is set to 4. First, when replacing the NL block with the simplified NL block and GC block, the performance can be regarded as on par and in top-1 accuracy, and in top-5 accuracy). As in COCO and ImageNet, adding more GC blocks further improves results and outperforms NL blocks with much less computation.

空洞研究。表10报告了空洞研究的结果。对于Kinetics实验，GC块的比例设置为4。首先，当用简化的NL块和GC块替换NL块时，性能可以认为与 和 在top-1精度上相当， 和 在top-5精度上相当。如在COCO和ImageNet中，增加更多的GC块进一步提高了结果，并且在计算量大大减少的情况下超过了NL块。

Comparison with Other Approaches. As shown in Table 11, we compare our approach with other state-of-the-art action recognition methods on Kinetics, and find that our GCNet outperforms GloRE [49] and NLNet [5].

与其他方法的比较。如表11所示，我们在Kinetics上与其他最新的动作识别方法进行了比较，发现我们的GCNet优于GloRE [49]和NLNet [5]。

# 5.4 Semantic Segmentation on Cityscapes

# 5.4 城市景观的语义分割

The Cityscapes [66] dataset is one of the most popular benchmarks for semantic segmentation, consisting of 5,000 high quality pixel-level finely annotated images and 20,000 coarsely annotated images captured from 50 different cities. Only the finely annotated part of the dataset is utilized in our experiments, and is divided into 2,975/500/1,525 for training, validation and testing. In total there are 30 semantic classes provided, 19 of which are used for evaluation. The standard mean Intersect over Union (mIoU) on the validation set is reported for measuring segmentation accuracy.

Cityscapes [66] 数据集是最流行的语义分割基准之一，包含来自50个不同城市的5,000张高质量像素级精细注释图像和20,000张粗略注释图像。在我们的实验中，只使用了数据集的精细注释部分，并分为2,975/500/1,525用于训练、验证和测试。总共提供了30个语义类别，其中19个用于评估。在验证集上报告标准的交并比（mIoU）以衡量分割精度。

The training setting and hyper-parameters strictly follow CC-Net [43]. The data are augmented by random scaling the original high resolution images by a factor in , then randomly cropping to patches. The poly learning policy is employed where the initial learning rate 0.01 is multiplied by . SGD training is performed on 4 GPUs with 2 images per GPU with Synchronized Batch Normalization for 160 epochs, which is roughly 60k steps. Following the practice of recent semantic segmentation approaches [24], [26], [41], [42], [43], ResNet-101 pretrained by [76], [77] is used as the backbone, where the downsampling operation in c4, c5 is removed and dilated convolution [78] is incorporated. The backbone is followed by a semantic segmentation head. Like the design in CCNet [43], the c5 feature is encoded by a context operator (e.g. CCNet, GCNet, SNLNet, NLNet) and concatenated with c5 before the pixel-wise classification layer. In the FCN [22] baseline, there is no context operator. As done in previous works [24], [41], [42], [43], an auxiliary head is added after the c4 stage output for a deep supervision loss. We use ratio for the GC block as default for semantic segmentation experiments.

训练设置和超参数严格遵循 CC-Net [43]。数据通过随机缩放原始 高分辨率图像的系数在 范围内，然后随机裁剪为 补丁来进行增强。采用了多项式学习策略，其中初始学习率 0.01 乘以 。在 4 个 GPU 上执行 SGD 训练，每个 GPU 处理 2 张图像，并使用同步批量归一化进行 160 个时期，大约是 60k 步。遵循最近语义分割方法的做法 [24]，[26]，[41]，[42]，[43]，使用 [76]，[77] 预训练的 ResNet-101 作为基础网络，其中移除了 c4, c5 中的下采样操作并集成了扩张卷积 [78]。基础网络后接一个语义分割头。与 CCNet [43] 中的设计类似，c5 特征由一个上下文操作符（例如 CCNet、GCNet、SNLNet、NLNet）编码，并与 c5 在像素级分类层之前连接。在 FCN [22] 基线中，没有上下文操作符。与之前的工作 [24]，[41]，[42]，[43] 一样，在 c4 阶段输出后添加了一个辅助头，用于深度监督损失。在语义分割实验中，我们使用默认的 作为 GC 块的比例。

Block Design. As shown in Table 12(a), the SNL head achieves performance comparable to the NL Head. Hence we argue that the accuracy gains by self-attention can be mainly ascribed the modeling of global context rather than the learning of pairwise relations. Moreover, all heads significantly boost the performance over the baseline, which indicates that long-range dependency is essential in the fine-grained semantic segmentation task. Note that with the GC block incorporated in the head, the GC blocks in the backbone do not have a significant effect because long range dependency is already exploited.

块设计。如表12(a)所示，SNL头部实现了与NL头部相当的性能。因此我们主张，自注意力带来的准确性增益主要归因于全局上下文的建模，而不是成对关系的学習。此外，所有头部在基线的基础上显著提升了性能，这表明在细粒度语义分割任务中长距离依赖是至关重要的。注意，当GC块被整合到头部时，骨干网络中的GC块没有显著效果，因为长距离依赖已经被利用。

Pooling and Fusion. The observations for the pooling and fusion in 12(b) are similar to those of object detection. Moreover, attention pooling with addition (GCNet) outperforms vanilla average pooling with scaling (SENet with layer norm) with almost the same #params and FLOPs. We conjecture that simply recalibrating channels does not effectively exploit rich semantic global context.

池化与融合。图12(b)中池化和融合的观察结果与目标检测中的相似。此外，带有加法的注意力池化（GCNet）在几乎相同的参数数量和浮点运算次数下，优于带有缩放的普通平均池化（带有层归一化的SENet）。我们推测，仅仅重新校准通道并不能有效地利用丰富的语义全局上下文。

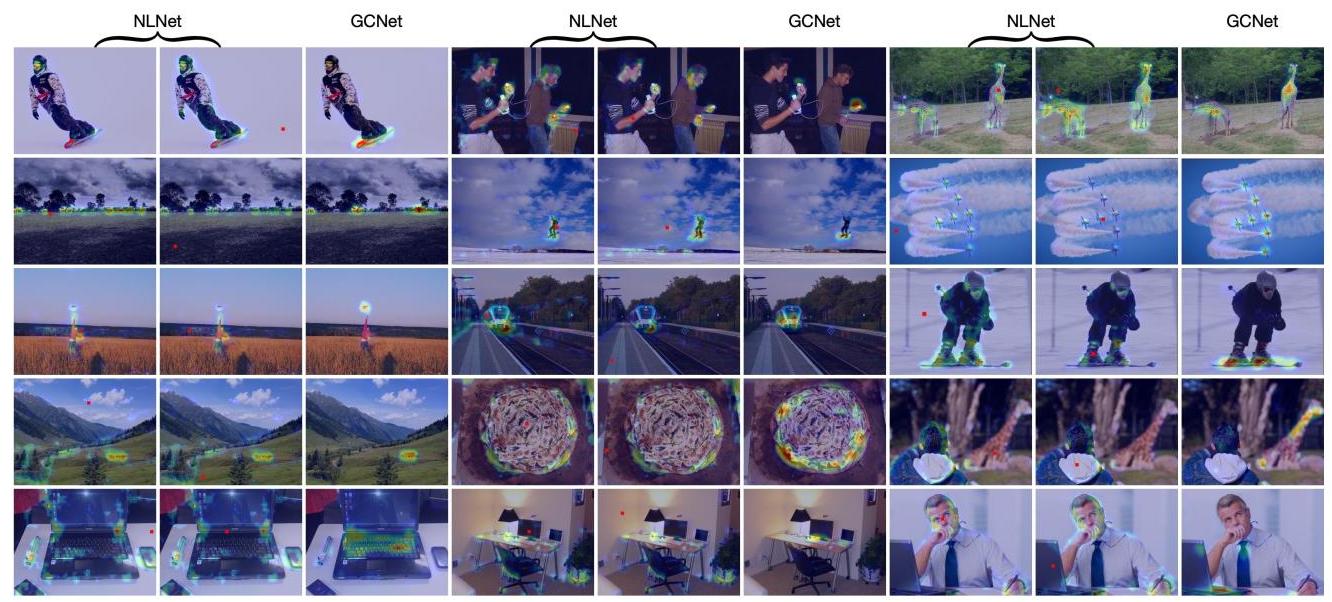


Fig. 5: Visualizations of context modeling attention maps (heatmaps) of GCNet and NLNet (red points denote query positions). Their learnt attention maps are mostly similar. Also, they learn to focus more on hard cases like relatively small size, deformation, occlusion, and blur. Best viewed in color.

图5：GCNet和NLNet上下文建模注意力图（热图）的可视化（红色点表示查询位置）。它们学到的注意力图大多相似。此外，它们学会更多地关注相对较小尺寸、形变、遮挡和模糊等困难案例。彩色查看效果最佳。

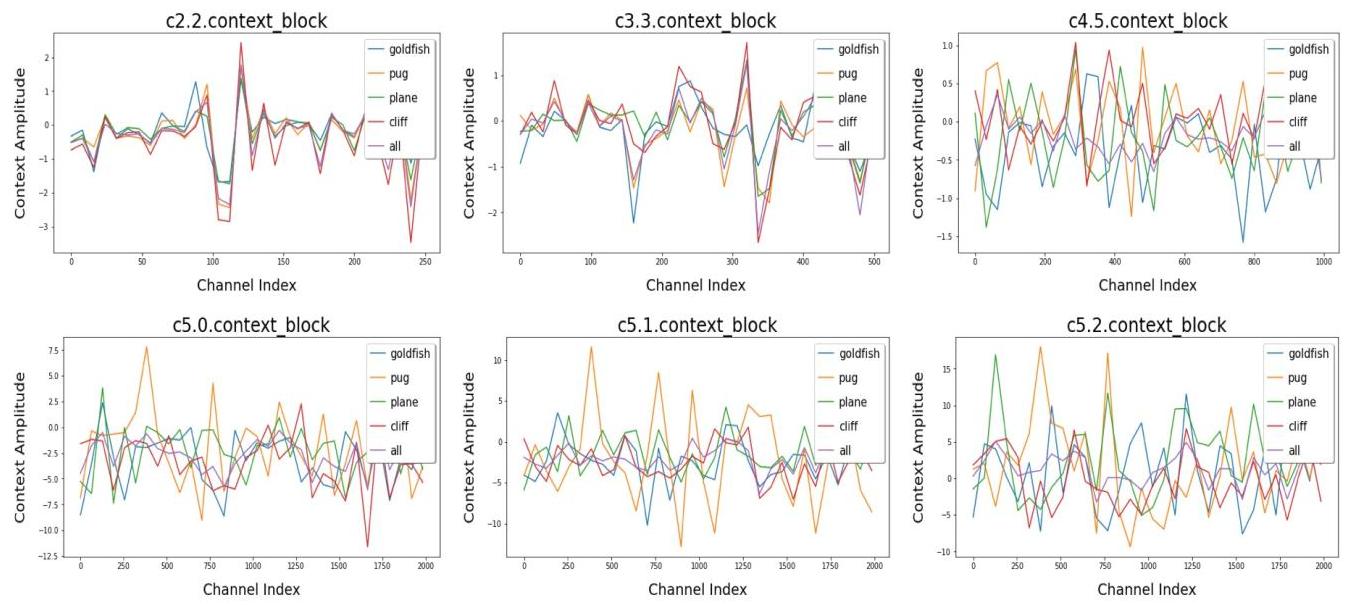


Fig. 6: Activation of output of the transform function at different stages of GCNet. Best viewed in color.

图6：GCNet中不同阶段的变换函数输出的激活。彩色查看效果最佳。

Comparison with Other Approaches. As shown in Table 13, we compare our approach with other state-of-the-art approaches on semantic segmentation of Cityscapes, and find that our GCNet achieves performance on par with DANet [42], ANN [79], CCNet [43] and NLNet [5].

与其他方法的比较。如表13所示，我们将我们的方法与Cityscapes语义分割的其他最先进方法进行了比较，并发现我们的GCNet实现了与DANet [42]、ANN [79]、CCNet [43]和NLNet [5]相当的性能。

# 5.5 Visualizations

# 5.5 可视化

Visualizations of Context Attention Map. In Figure 5, we randomly choose fifteen images from the COCO dataset and visualize their attention maps (softmax output of context modeling module) for GCNet and NLNet. We can observe that NLNet learns similar attention maps for different query points in most cases, which are also similar to the attention maps learnt by GCNet. In addition, we observe that the two models usually focus on small or thin objects like frisbee, skateboard, and snowboard. This may facilitate the detection of these objects, and the accuracy is hence boosted. Also note that the human body is an exception, which is less attended. We hypothesize the reason is because the person class is common enough in the COCO dataset and it is not hard to be detected.

上下文注意力图的可视化。在图5中，我们从COCO数据集中随机选择了十五张图像，并可视化了它们在GCNet和NLNet上的注意力图（上下文建模模块的softmax输出）。我们可以观察到，在大多数情况下，NLNet为不同的查询点学习到的注意力图相似，这些图也与GCNet学习到的注意力图相似。此外，我们还观察到两个模型通常会关注飞盘、滑板和滑雪板等小型或细长物体。这可能会促进这些物体的检测，并因此提高准确率。请注意，人体是一个例外，其关注度较低。我们推测原因是COCO数据集中人物类别足够常见，且易于检测。

Output Activations of GC Block. We follow [2] to visualize the output activations of GC blocks in different layers. As depicted

GC块输出激活。我们遵循[2]的方法，可视化不同层中GC块的输出激活。

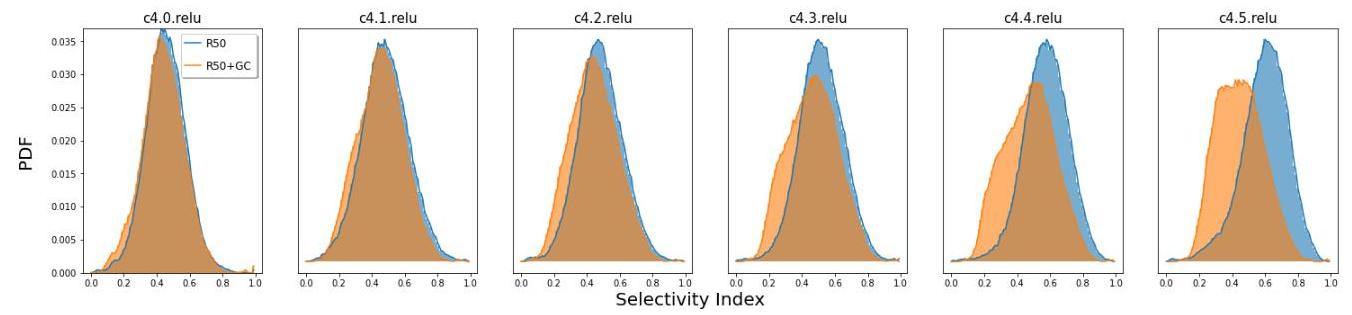


Fig. 7: Distributions of class selectivity index of ResNet-50 baseline and ResNet-50+GCNet on different layers. For deeper layers, GCNet exhibits less class selectivity compared to the baseline. Best viewed in color.

图7：ResNet-50基线和ResNet-50+GCNet在不同层上的类别选择性指数分布。对于更深的层，与基线相比，GCNet表现出较少的类别选择性。彩色查看效果最佳。

| (a) Block Design | | | |
| --- | --- | --- | --- |
|  | #params(M) | FLOPs(G) |  |
| baseline | 70.96 | 646.88 | 75.42 |
| NL Head | 71.22 | 697.16 | 77.59 |
| SNL Head | 71.22 | 646.89 | 77.22 |
| GC head | 71.09 | 646.89 | 78.55 |
| GC backbone (c3-c5) | 89.92 | 647.19 | 78.49 |
| GC backbone + GC Head | 90.05 | 647.20 | 78.67 |

(b) Pooling and Fusion

(b) 池化和融合

|  | #params(M) | FLOPs(G) |  |
| --- | --- | --- | --- |
| baseline | 70.96 | 646.88 | 75.42 |
| avg+scale | 71.09 | 646.89 | 76.86 |
| avg+add | 71.09 | 646.89 | 77.84 |
| att+scale | 71.09 | 646.89 | 77.49 |
| att+add | 71.09 | 646.89 | 78.55 |

TABLE 12: Ablation study of GCNet with ResNet-101 on semantic segmentation on Cityscapes validation set.

表12：在Cityscapes验证集上进行语义分割时，GCNet与ResNet-101的消融研究。

|  | #params(M) | FLOPs(G) |  |
| --- | --- | --- | --- |
| baseline | 70.96 | 646.88 | 75.52 |
| DANet [42] | 71.29 | 709.18 | 79.88 |
| ANN | 67.89 | 632.10 | 79.32 |
| [43] | 71.49 | 653.52 | 78.90 |
| NLNet [5] | 71.22 | 697.16 | 78.57 |
|  | 71.09 | 646.89 | 78.95 |

TABLE 13: Comparison of state-of-the-art methods with ResNet- 101 on semantic segmentation with stronger augmentation on Cityscapes validation set. The methods denoted with " " marker produce the pixel-wise classification logits by the concatenation of the stride- backbone and the context head followed by a convolution layer, while the others directly utilize the context head features without concatenating the 2048-dim c5 features.

表13：在Cityscapes验证集上，使用更强的增强方法进行语义分割时，当前最先进方法与ResNet-101的比较。标记为 " " 的方法通过拼接步长为 的主干网络和上下文头部，然后接一个 卷积层来产生像素级的分类logits，而其他方法则直接使用上下文头部特征，而不拼接2048维的c5特征。

in Figure 6, the channel activations are class-agnostic in the shallow layers and more class-dependent in the deeper layers. It is intuitive since for neurons closer to the final classification layer, a higher correlation between activation and class label is expected.

在图6中，浅层通道激活对于类别是不可区分的，而在更深层则更依赖于类别。这是直观的，因为对于接近最终分类层的神经元，期望激活与类别标签之间有更高的相关性。

Illustration of Class Selectivity. We use the class selectivity index proposed in [80] to study the effect of global context modeling on learned representations. In Figure 7, we plot the distribution of the class selectivity index on ImageNet. We use the last activation of each block in the c4 stage to compute the class selectivity index. The observed pattern is similar to that in GENet [45]. The distributions are almost the same in the first blocks. As the depth increases, GCNet begins to diverge from the baseline. And as shown in the last plot (c4.5.relu) in Figure 7, GCNet exhibits much less class selectivity. Also pointed out in [45], we speculate that there are some cases that suffer from local ambiguity, which would push the baseline network to specialize some neurons to overcome it. Note that the global context computed by GCNet may avoid this burden thus resulting in less class selectivity.

类选择性的说明。我们使用文献[80]中提出的类选择性指数来研究全局上下文建模对学习表征的影响。在图7中，我们绘制了ImageNet上类选择性指数的分布。我们使用c4阶段每个块的最后激活来计算类选择性指数。观察到的模式与GENet [45]中的相似。在最初的块中，分布几乎相同。随着深度的增加，GCNet开始与基线分离。如图7中的最后一个图（c4.5.relu）所示，GCNet显示出更少的类选择性。正如[45]中指出的，我们推测有些情况受到局部模糊的影响，这会推动基线网络使一些神经元专门化以克服它。请注意，GCNet计算的全局上下文可能会避免这种负担，从而产生更少的类选择性。

# 6 Conclusion

# 6 结论

The long-range dependency modeling of non-local networks intends to model query-specific global context, but we have found empirically that it only models query-independent context on several important visual recognition tasks. Based on this, we simplify non-local networks and abstract this simplified version to a global context modeling framework. Then we propose a novel instantiation of this framework, the GC block, which is lightweight and can effectively model long-range dependency. Our GCNet is constructed via applying GC blocks to multiple layers, which generally outperforms simplified NLNet on major benchmarks for various recognition tasks.

非局部网络的长距离依赖建模旨在模拟特定于查询的全局上下文，但我们通过经验发现在几个重要的视觉识别任务上，它仅模拟独立于查询的上下文。基于这一点，我们简化了非局部网络，并将这种简化版本抽象为一个全局上下文建模框架。然后我们提出了这个框架的一个新颖实例化，即GC块，它轻量级且能有效模拟长距离依赖。我们的GCNet通过在多层应用GC块构建而成，通常在多种识别任务的主要基准上优于简化的NLNet。

We have verified that the global context block can benefit multiple visual recognition tasks. In the future, the global context block may be extended to the generative models [81], [82], graph learning models [56], [83], and self-supervised models [84].

我们已经验证了全局上下文块可以惠及多种视觉识别任务。在未来，全局上下文块可能会扩展到生成模型 [81], [82]，图学习模型 [56], [83]，以及自监督模型 [84]。

# References

# 参考文献

[1] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Genet: Non-local networks meet squeeze-excitation networks and beyond," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019.

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[3] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588-3597.

[4] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151-7160.

[5] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

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

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," arXiv preprint arXiv:1812.03982, 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

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

[13] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700-4708.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492-1500.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[17] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764-773.

[18] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," arXiv preprint arXiv:1811.11168, 2018.

[19] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in International Conference on Computer Vision, 2019.

[20] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8697- 8710.

[21] Z. Zhong, J. Yan, W. Wu, J. Shao, and C.-L. Liu, "Practical block-wise neural network architecture generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2423-2432.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440.

[23] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint arXiv:1606.02147, 2016.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881-2890.

[25] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 405-420.

[26] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801-818.

[28] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in Advances in neural information processing systems, 2016, pp. 379-387.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-

[30] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263-7271.

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

[32] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 305-321.

[33] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933- 1941.

[34] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510-4520.

[37] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848-6856.

[38] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116-131.

[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299-6308.

[40] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533-5541.

[41] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," arXiv preprint arXiv:1809.00916, 2018.

[42] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146-3154.

[43] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Cc-net: Criss-cross attention for semantic segmentation," arXiv preprint arXiv:1811.11721, 2018.

[44] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in Proceedings of the European conference on computer vision (ECCV), 2020.

[45] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in Advances in Neural Information Processing Systems, 2018, pp. 9423-9433.

[46] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267-283.

[47] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19.

[48] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin, "A convolutional encoder model for neural machine translation," arXiv preprint arXiv:1611.02344, 2016.

[49] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 433- 442.

[50] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina et al., "State-of-the-art speech recognition with sequence-to-sequence models," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4774-4778.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[52] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," arXiv preprint arXiv:1904.05873, 2019.

[53] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," arXiv preprint arXiv:1905.03197, 2019.

[54] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," arXiv preprint arXiv:1711.04043, 2017.

[55] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 1243-1252.

[56] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Ben-gio, "Graph attention networks," arXiv preprint arXiv:1710.10903, 2017.

[57] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," arXiv preprint arXiv:1805.08318, 2018.

[58] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156-3164.

[59] J. Xu, Y. Cao, Z. Zhang, and H. Hu, "Spatial-temporal relation networks for multi-object tracking," in International Conference on Computer Vision, 2019.

[60] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," 2020.

[61] J. Gu, H. Hu, L. Wang, Y. Wei, and J. Dai, "Learning region features for object detection," in Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[62] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," CoRR, vol. abs/1311.2901, 2013. [Online]. Available:|http://arxiv.org/abs/1311.2901

[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740-755.

[64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248-255.

[65] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijaya-narasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[66] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213-3223.

[67] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

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

[69] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91-99.

[71] F. Massa and R. Girshick, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch," https://github.com/facebookresearch/ maskrcnn-benchmark 2018, accessed: [2019.03.22].

[72] Y. Wu and K. He, "Group normalization," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19.

[73] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," arXiv preprint arXiv:1811.08883, 2018.

[74] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154-6162.

[75] J. Xie, T. He, Z. Zhang, H. Zhang, Z. Zhang, and M. Li, "Bag of tricks for image classification with convolutional neural networks," arXiv preprint arXiv:1812.01187, 2018.

[76] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[77] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Tor-ralba, "Semantic understanding of scenes through the ade20k dataset," International Journal on Computer Vision, 2018.

[78] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[79] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 593-602.

[80] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," arXiv preprint arXiv:1803.06959, 2018.

[81] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672- 2680.

[82] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[83] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[84] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2566-2576.