SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2023, Vol. 66 142105:1–142105:19
https://doi.org/10.1007/s11432-021-3590-1

Augmented FCN: rethinking context modeling for semantic segmentation

Dong ZHANG[1], Liyan ZHANG[2] & Jinhui TANG[1*]

[1]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing* 210094*, China;*
[2]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106*, China*

Received 26 December 2021/Revised 8 June 2022/Accepted 28 July 2022/Published online 9 February 2023

Abstract The effectiveness of modeling contextual information has been empirically shown in numerous computer vision tasks. In this paper, we propose a simple yet efficient augmented fully convolutional network (AugFCN) by aggregating content- and position-based object contexts for semantic segmentation. Specifically, motivated because each deep feature map is a global, class-wise representation of the input, we first propose an augmented nonlocal interaction (AugNI) to aggregate the global content-based contexts through all feature map interactions. Compared to classical position-wise approaches, AugNI is more efficient. Moreover, to eliminate permutation equivariance and maintain translation equivariance, a learnable, relative position embedding branch is then supportably installed in AugNI to capture the global position-based contexts. AugFCN is built on a fully convolutional network as the backbone by deploying AugNI before the segmentation head network. Experimental results on two challenging benchmarks verify that AugFCN can achieve a competitive 45.38% mIoU (standard mean intersection over union) and 81.9% mIoU on the ADE20K val set and Cityscapes test set, respectively, with little computational overhead. Additionally, the results of the joint implementation of AugNI and existing context modeling schemes show that AugFCN leads to continuous segmentation improvements in state-of-the-art context modeling. We finally achieve a top performance of 45.43% mIoU on the ADE20K val set and 83.0% mIoU on the Cityscapes test set.

Keywords semantic segmentation, context modeling, long-range dependencies, attention mechanism

Citation Zhang D, Zhang L Y, Tang J H. Augmented FCN: rethinking context modeling for semantic segmentation. Sci China Inf Sci, 2023, 66(4): 142105, https://doi.org/10.1007/s11432-021-3590-1

1 Introduction

Semantic segmentation (SS) aims to assign a unique class label to each pixel of a given image, which is one of the most fundamental yet challenging topics in computer vision research. Over the years, SS has been intensively studied and applied to a wide range of potential applications, e.g., computer-aided diagnosis [1], autonomous driving [2], and augmented reality [3].

Recently, the progressive SS methods have mainly been based on fully convolutional networks (FCNs) [4]. However, because of the fixed, local receptive fields and short-range dependencies, the deep feature maps of FCN-based methods can only aggregate the local object contexts, which are insufficient for some complicated scenarios [5–9]. For this reason, extraordinary progress has been made in modeling global contexts. In general, the existing schemes can be divided into category-I atrous convolution [10] or pooling layer-based methods and category-II self-attention [11] or nonlocal interaction [12]-based methods. Specifically, for the elementary category-I, global object contexts are passively captured by enlarging effective receptive fields. One approach is to use atrous convolutions [10] to replace downsampling layers in the backbone network [13–15] such that the receptive fields can be enlarged while a high resolution of output feature maps is maintained. The other approach is to stack pyramidal atrous convolutions or pooling layers in the head network [5, 16–19], e.g., PPM in PSPNet [16], ASPP in DeepLab schemes [13, 18, 19], and MPM in SPNet [5]. However, the category-I methods have the disadvantage of only capturing contexts within square regions. In particular, for some slender objects (e.g., "`rail track`", "`pole group`",

* Corresponding author (email: jinhuitang@njust.edu.cn)

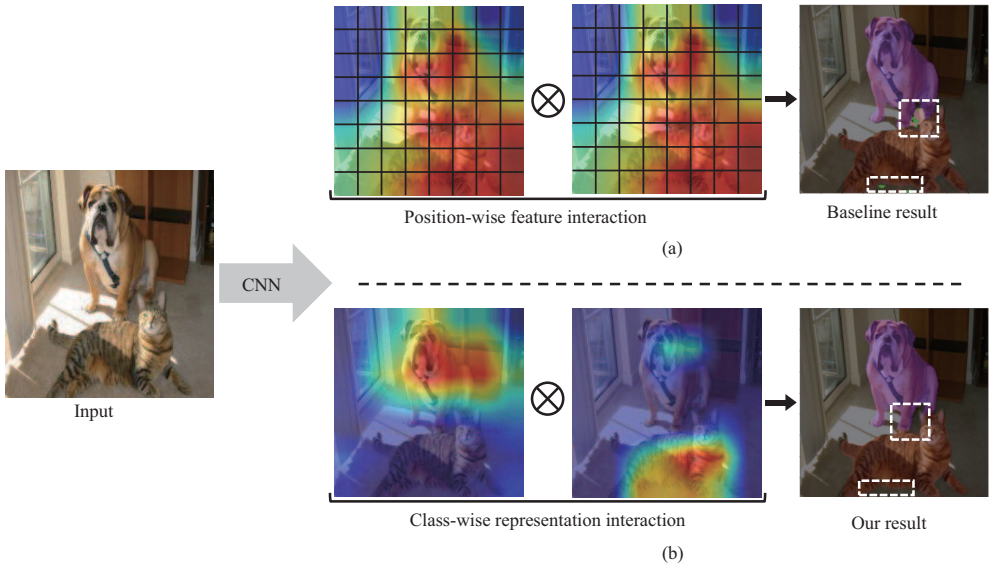


Figure 1 (Color online) An illustration of context modeling schemes. (a) The position-wise context modeling approach, where every feature grid is taken into consideration for calculating a global context mapping matrix. (b) Our proposed context modeling scheme — augmented nonlocal interaction (AugNI), where object contexts are obtained by class-wise representation interactions. In AugNI, the problem of large computational overhead of the position-wise context modeling approach can be alleviated. Just for simplicity, we do not distinguish between content-based and position-based contexts in this figure. The white dashed frames in the last column highlight the improved areas predicted by our model. Best viewed in color.

and "`traffic sign`" in Cityscapes [20] dataset), methods in the category-I reservoir may introduce irrelevant and even noisy surrounding information. The other obvious disadvantage is that modules based on atrous convolutions and pooling layers are inherently sub-sampling-based methods, which may cause detailed spatial information loss. For the progressive category-II, as illustrated in Figure 1(a), global object contexts are actively aggregated through long-range dependencies by position-wise feature interactions [8, 9, 21–27]. However, despite the elementary success in SS, because a global context mapping must be computed for every feature position, self-attention and nonlocal operation virtually bring much computational overhead in time and space [21, 23, 28]. Moreover, self-attention and nonlocal operation have another noteworthy disadvantage of being only content-based methods [11, 12] without having relative position information, which makes them permutation equivariant [29–31] and thus ineffective on highly structured data (e.g., music, images, and videos), as empirically shown in numerous studies [32–34].

At this point, context modeling for SS should be further developed. We start with a novel thought that is also our key motivation. For context modeling, is establishing a global context mapping really necessary for every feature position? To answer such a question, we start with an interpretation of the deep feature maps involved in SS. Suppose we have a set of deep feature maps $\boldsymbol{X} \in \mathbb{R}^{h\times w\times c}$ obtained from a trained FCN, where h, w, and c denote the height, width, and channel dimension, respectively. From the concept of "semantics" in vision tasks, each feature map $\boldsymbol{x}_i \in \mathbb{R}^{h\times w}$ can be considered a global, class-wise representation of the input image and does not correspond to any position-wise matter; i.e., all feature positions within $\boldsymbol{x}_i$ mainly respond to one category of objects [35–37]. Under this consideration, if we could use the entire $\boldsymbol{x}_i$ to model the global object contexts, as illustrated in Figure 1(b), then establishing a global context mapping for every feature position can be avoided. Therefore, the problem of the large computational overhead of the position-wise context modeling schemes (e.g., self-attention [11] and nonlocal operation [12]) can be alleviated. Moreover, contexts obtained in such a concise way also meet the definition of global object contexts, i.e., relations between different objects and between objects and the background.

In this paper, as illustrated in Figure 2, we propose a simple yet efficient global context modeling scheme, termed augmented nonlocal interaction (AugNI), to aggregate content- and position-based object contexts. Specifically, motivated by each feature map being a global, class-wise image representation, we first use AugNI to compute a semantic affinity matrix (SAM) through the standard dot-product operation based on all feature maps, where each entry in the SAM represents similarities between one class-wise representation and all other class-wise representations. Compared to the classical position-wise context mapping methods in self-attention and nonlocal operation, the SAM has higher efficiency and less

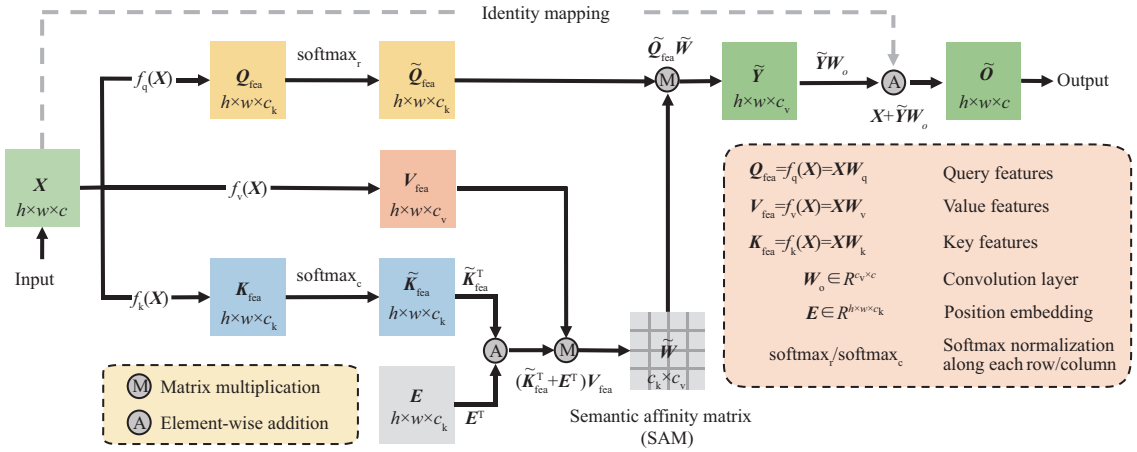


Figure 2 (Color online) An illustration of our proposed AugNI. The input is a set of deep feature maps $\boldsymbol{X} \in \mathbb{R}^{h\times w\times c}$, and output $\tilde{\boldsymbol{O}} \in \mathbb{R}^{h\times w\times c}$ has the same scale of the input but with more contextual information. The entire feature maps (i.e., $\boldsymbol{Q}_{\rm fea}$, $\boldsymbol{K}_{\rm fea}$, and $\boldsymbol{V}_{\rm fea}$) are used for modeling the global object contexts within AugNI, rather than establishing a global context mapping for every feature position. Therefore, compared to the classical nonlocal operation [12], our AugNI has less computational overhead and contains the beneficial relative position information, which makes AugNI suitable for images.

computational overhead (see Subsection 3.2). Then, we weighted the sum of each entry in the SAM for the corresponding class-wise representation to aggregate the content based on global contexts. Furthermore, to eliminate the permutation equivariance and maintain the translation equivariance of AugNI, a learnable relative position embedding branch is paralleled and installed in the SAM to capture the global, position-based object contexts. Finally, the AugNI output is the summation of an identity mapping branch from the input and a set of processed feature maps containing sufficient contexts. AugNI has the same spirit of self-attention and nonlocal operation, but it is more suited for SS tasks (see Subsection 4.3).

To demonstrate its superiority, we deployed AugNI on a fully convolutional network as the backbone, named augmented FCN (AugFCN), for SS tasks in an end-to-end training fashion. The experimental results show that AugFCN with ResNet-101 [38] can achieve a competitive performance on two challenging benchmarks of 45.38% on the ADE20K [39] val set and 81.9% on the Cityscapes [20] test set. In terms of model efficiency, compared to the standard self-attention [11] and nonlocal operation [12], AugFCN with ResNet-50 [38] can reduce at most 48.0 M parameters and 3.15 GFLOPs (see Subsection 4.3.2). Moreover, we also validated the effectiveness of the joint implementation of AugFCN and the existing context modeling schemes. The results show that AugFCN can lead to continuous improvements in state-of-the-art context modeling methods. We finally achieve a top performance of 45.43% on the ADE20K val set and 83.0% on the Cityscapes test set.

The main contributions of this work are summarized as follows.

- We proposed an effective yet efficient context modeling scheme that can aggregate the content- and position-based object contexts.
- We demonstrated the effectiveness of AugFCN as well as the joint implementation of AugFCN and the existing context modeling methods.
- We demonstrated the efficiency of AugFCN on the standard mean intersection over union (mIoU) with less computational overhead than most existing context modeling methods.
- The experimental results showed that we finally achieved the top performance on two challenging benchmarks, ADE20K and Cityscapes.

2 Related work

2.1 Semantic segmentation

Semantic segmentation is one of three fundamental tasks (i.e., image classification, object detection, and semantic segmentation) in the computer vision domain, which has been extensively studied in the recent past [40, 41]. FCN [4] is a classical approach that uses a fully convolutional network in modern semantic segmentation systems with a deep artificial neural network (e.g., VGG [42], ResNet [38], and DenseNet [43]) as its backbone. Most of the subsequent semantic segmentation models [5,8,9,13,17,21,44]

stand on the shoulders of FCN. A pivotal challenge for semantic segmentation is the resolution reduction of the deep output feature maps. To maintain a high-resolution for the output feature maps, the existing methods either use an encoder-decoder architecture [45–49] to gradually fuse low-resolution deep feature maps with high-resolution shallow ones, e.g., U-Net [48], Encoder-decoder [19], and the De-convolution Network [49], or they use the traditional convolution architecture, where atrous convolutions [10] are used in the backbone, e.g., DeepLab schemes [13, 18, 19], CFNet [9], and SPNet [5]. In our work, we also use the fully convolutional architecture as the backbone. In particular, following [5, 9, 23], we use classical ResNet [38] with atrous convolutions on block3 and block4 such that the resolution of the output feature maps is 1/8 of the input.

2.2 Context modeling

The effectiveness and importance of capturing object contexts have been empirically shown over a wide range of computer vision tasks [11,12,22,50,51], e.g., onfocus detection [52], object detection [53], and object localization [54]. These methods enhance feature contexts by increasing the effective receptive fields. For semantic segmentation, DeepLab schemes [13,18,19] and PSPNet [16] open a new era for modeling the global object contexts through a multi-scale feature pyramid. For example, the pyramid pooling module in PSPNet uses four global pooling layers with different downsampling rates to establish a segmentation feature pyramid such that four scales of the object contexts can be captured. Afterward, the following methods, e.g., Dense-ASPP [55], SPNet [5], and CGNet [56], continue to use this framework. This type of context modeling method is advantageous because the computational overhead is relatively small, but the captured contexts are limited to the square regions, which are unfriendly to some slender objects [5, 23]. Another mainstream approach to aggregating object contexts is to use the self-attention mechanism [11] or nonlocal operation [12]. The representative methods are CFNet [9], ANNN [21], DANet [44], and CCNet [57]. However, context modeling methods of this type always suffer from the disadvantages of having a large computational overhead and only considering content-based interactions [29,31,58], which are not suitable for image data. In our work, we also focus on context modeling and propose an effective yet efficient context modeling module to aggregate content- and position-based object contexts. Compared to the previous context modeling methods, our method not only has a higher efficiency but also can encode the relative position information. Extensive experimental results on two challenging benchmarks demonstrated its effectiveness on images.

2.3 Attention learning

To make the neural network heed more important regions, the attention mechanism was designed and has brought benefits to many natural language processing and computer vision tasks. SCA-CNN [59] and SENet [60] are two inchoate methods for using the channel attention mechanism to enhance model representation ability. Moreover, SKNet [61] and ResNeSt [62] propose more effective channel attention mechanisms for computer vision tasks by splitting and combining feature maps. Nonlocal operation [12] has a similar architecture as self-attention [11], but it is proposed for capturing the focused areas through a large attention map. In the semantic segmentation domain, DANet [44], ANNN [21], and CFNet [9] use the self-attention and nonlocal operation to aggregate the global object contexts through position-wise long-range dependencies. To alleviate the problem of large computational overhead, CCNet [23] was recently proposed to learn the long-range dependencies with feature positions in the same row and column. In this work, our implementation is also inspired by the progressive attention learning mechanisms: efficient attention [35] and Lambda networks [63]. Both of these learning mechanisms aim to improve attention learning mechanism efficiency. However, in contrast to these two approaches, our proposed method has different viewpoints, motivations, and application scenarios. By observing the existing deep feature maps for SS, we treat each deep feature map as a global, class-wise representation and use the entire feature map to aggregate the global object contexts. Therefore, our method has the advantage of high efficiency. Furthermore, our method also considers the relative position information and contains a learnable relative position embedding branch, which has been empirically shown to benefit the semantic segmentation task.

3 Our approach

In this section, we show implementation details of the proposed AugFCN for semantic segmentation by aggregating content- and position-based object contexts. We first revisit the classical position-wise nonlocal operation[1)] on images (in Subsection 3.1). Then, the proposed AugNI is introduced (in Subsection 3.2). Finally, we introduce AugFCN, which is built on the standard FCN with ResNet as the backbone, by deploying AugNI on the head network (in Subsection 3.3).

3.1 Revisiting nonlocal interaction

For an input RGB image, we can obtain a set of deep feature maps $\boldsymbol{X} \in \mathbb{R}^{h\times w\times c}$ through a trained FCN, where h, w, and c denote the height, width, and channel dimension, respectively. The classical nonlocal operation [12] can be used to actively capture the global object contexts through long-range dependencies by position-wise feature interactions.

Specifically, the nonlocal operation first uses three linear functions, i.e., $f_{\mathrm{q}}(\cdot)$, $f_{\mathrm{k}}(\cdot)$, and $f_{\mathrm{v}}(\cdot)$, to map feature maps $\boldsymbol{X}$ into query features $\boldsymbol{Q}_{\mathrm{fea}} = f_{\mathrm{q}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\mathrm{q}}$, key features $\boldsymbol{K}_{\mathrm{fea}} = f_{\mathrm{k}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\mathrm{k}}$, and value features $\boldsymbol{V}_{\mathrm{fea}} = f_{\mathrm{v}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\mathrm{v}}$, respectively. $\boldsymbol{W}_{\mathrm{q}}, \boldsymbol{W}_{\mathrm{k}} \in \mathbb{R}^{c\times c_{\mathrm{k}}}$, and $\boldsymbol{W}_{\mathrm{v}} \in \mathbb{R}^{c\times c_{\mathrm{v}}}$ are three learnable linear transformations. Without loss of generality, we assume that $c_{\mathrm{k}} = c_{\mathrm{v}} < c$. Then a global context mapping matrix $\boldsymbol{W} \in \mathbb{R}^{hw\times hw}$ between $\boldsymbol{Q}_{\mathrm{fea}}$ and $\boldsymbol{K}_{\mathrm{fea}}$ can be calculated using the following procedure:

$$\boldsymbol{W} = \mathrm{softmax}\left((\boldsymbol{X}\boldsymbol{W}_{\mathrm{q}})(\boldsymbol{X}\boldsymbol{W}_{\mathrm{k}})^{\mathrm{T}}\right), \tag{1}$$

where each item w_{pq} in $\boldsymbol{W}$ denotes the similarity between the p-th query feature $\boldsymbol{q}^{\mathrm{p}}_{\mathrm{fea}} \in \mathbb{R}^{1\times c}$ and the q-th key feature $\boldsymbol{k}^{\mathrm{q}}_{\mathrm{fea}} \in \mathbb{R}^{1\times c}$. Because every feature position of $\boldsymbol{Q}_{\mathrm{fea}}$ and $\boldsymbol{K}_{\mathrm{fea}}$ participates in calculating $\boldsymbol{W}$, we refer to this method as the position-wise interaction (the same thing happens in self-attention [11]). Softmax$(\cdot)$ is used for feature normalization along each row. On the basis of this similarity, the global object contexts $\boldsymbol{Y}$ can be obtained through a weighted sum operation with value features $\boldsymbol{V}_{\mathrm{fea}}$ as follows:

$$\boldsymbol{Y} = \boldsymbol{W}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{v}}). \tag{2}$$

Finally, the output $\boldsymbol{O}$ of the nonlocal operation is an element-wise summation between the input feature maps $\boldsymbol{X}$ and the processed feature maps with global object contexts $\boldsymbol{Y}$:

$$\boldsymbol{O} = \boldsymbol{X} + \boldsymbol{Y}\boldsymbol{W}_{\mathrm{o}}, \tag{3}$$

where $\boldsymbol{W}_{\mathrm{o}} \in \mathbb{R}^{c_{\mathrm{v}}\times c}$ denotes a 1×1 convolution layer, which is used to increase the channel dimension of $\boldsymbol{Y}$ into c.

Multi-head nonlocal. Instead of mapping feature maps $\boldsymbol{X}$ into a uniform space, it has been empirically shown that a more effective approach is to project $\boldsymbol{X}$ into d head subspaces and then concatenate the global object contexts obtained in each head as the final object contexts [9,11,29,64]. This procedure can be expressed as follows:

$$\boldsymbol{O}_{\mathrm{mn}} = f_{\mathrm{c}}(\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_{\mathrm{d}})\boldsymbol{W}_{\mathrm{mn}}, \tag{4}$$

where $\boldsymbol{O}_{\mathrm{mn}}$ is the final output of the multi-head nonlocal interaction. $f_{\mathrm{c}}(\cdot)$ denotes the feature map concatenation operation along the channel dimension, and $\boldsymbol{W}_{\mathrm{mn}} \in \mathbb{R}^{c\times c}$ denotes a learnable linear transformation for feature fusion.

Although nonlocal operations and self-attention have achieved satisfactory success in semantic segmentation [8,9,21,24], they suffer from the apparent disadvantage of a large computational overhead in time and space. For example, the computational complexity of generating a single context mapping matrix $\boldsymbol{W}$ (i.e., a single-head nonlocal operation) is $\mathcal{O}(h^2w^2)$. This phenomenon will be worse if the resolution of $\boldsymbol{X}$ is extremely high (e.g., using atrous convolutions [10] in the backbone network) or if we use a multi-head nonlocal operation whose computational complexity is as high as $\mathcal{O}(dh^2w^2)$. Another noteworthy disadvantage is the nonlocal operation and self-attention being only content-based interactions, which are not favorable for highly structured image data [23,29,31].

1) Because self-attention and nonlocal operation on images are deployed identically, we only introduce the nonlocal operation [12] for simplicity.

3.2 AugNI

In this work, as illustrated in Figure 2, we propose an effective yet efficient context modeling scheme for semantic segmentation, named AugNI, to aggregate content- and position-based object contexts. AugNI inputs a set of deep feature maps $\boldsymbol{X}$ and outputs a set of enhanced feature maps $\tilde{\boldsymbol{O}}$, which have the same scale as $\boldsymbol{X}$ but with more contextual and relative position information.

Motivation. From the common concept of "semantics" in the computer vision domain [65, 66], each feature map $\boldsymbol{x}_i$ (derived from a group of deep feature maps $\boldsymbol{X}$) can be considered a global, class-wise feature representation of the input image and does not correspond to any position-wise matter; i.e., all feature positions within $\boldsymbol{x}_i$ mainly respond to only one category of objects. In particular, this phenomenon becomes more pronounced as the depth of the feature map increases [35–37, 67]. For example, for the $\boldsymbol{X}$ of an image with a person riding a bicycle, one feature map may correspond to the "`person`", one may correspond to the "`bicycle`", and another one may correspond to the "`background`". In addition, recall that the general meaning of object contexts for semantic segmentation is the relation between different categories of objects and between foreground objects and the background. Therefore, we can use the entire feature map (i.e., each class-wise image representation) to calculate object contexts rather than feature positions, as in the nonlocal operation [12] and self-attention [11]. On the basis of this motivation, the problem of large computational overhead can be alleviated. Below, we introduce our method in detail.

Base AugNI. AugNI mainly comprises two components: the global content-based and global position-based object contexts. To satisfy the definition of "similarity" and simplify the calculation procedure, AugNI first normalizes the key features $\boldsymbol{K}_{\rm fea}$ into $\tilde{\boldsymbol{K}}_{\rm fea}$ through a standard softmax function along each column. Then, an SAM $\tilde{\boldsymbol{W}} \in \mathbb{R}^{c_{\rm k} \times c_{\rm v}}$ can be obtained using

$$\tilde{\boldsymbol{W}} = (\tilde{\boldsymbol{K}}_{\rm fea} + \boldsymbol{E})^{\rm T} \boldsymbol{V}_{\rm fea} = \underbrace{\tilde{\boldsymbol{K}}_{\rm fea}^{\rm T}(\boldsymbol{X}\boldsymbol{W}_{\rm v})}_{\text{content SAM}} + \underbrace{\boldsymbol{E}^{\rm T}(\boldsymbol{X}\boldsymbol{W}_{\rm v})}_{\text{position SAM}}, \tag{5}$$

where $\boldsymbol{E} \in \mathbb{R}^{h \times w \times c_{\rm k}}$ denotes a learnable embedding tensor, which is used to encode the relative position information of each class-wise representation. $\tilde{\boldsymbol{K}}_{\rm fea}^{\rm T}(\boldsymbol{X}\boldsymbol{W}_{\rm v})$ and $\boldsymbol{E}^{\rm T}(\boldsymbol{X}\boldsymbol{W}_{\rm v})$ denote the content-based and position-based SAMs, respectively. Each entry $\tilde{w}_{ij}$ in the content-based SAM follows the same definition as inefficient attention [35] and nonlocal operation [12], which denotes the similarity between the i-th feature map $\boldsymbol{k}_{\rm fea}^i \in \mathbb{R}^{hw \times 1}$ in $\boldsymbol{K}_{\rm fea}$ (corresponding to a class-wise representation) and the other j-th feature map $\boldsymbol{v}_{\rm fea}^j \in \mathbb{R}^{hw \times 1}$ in $\boldsymbol{V}_{\rm fea}$ (corresponding to the other class-wise representations). Conversely, each entry $\tilde{w}_{ij}$ in the position-based SAM denotes the relative position embedding between the i-th learnable position embedding vector $\boldsymbol{e}_i \in \mathbb{R}^{hw \times 1}$ in $\boldsymbol{E}$ and the other j-th feature map $\boldsymbol{v}_{\rm fea}^j$ in $\boldsymbol{V}_{\rm fea}$. Element-wise summation between the content-based SAM and position-based SAM forms the final SAM output, which contains the content- and relative position-based information. On the basis of this information, our global object contexts can be calculated via

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Q}}_{\rm fea}\tilde{\boldsymbol{W}} = \underbrace{\tilde{\boldsymbol{Q}}_{\rm fea}\left(\tilde{\boldsymbol{K}}_{\rm fea}^{\rm T}(\boldsymbol{X}\boldsymbol{W}_{\rm v})\right)}_{\text{content contexts}} + \underbrace{\tilde{\boldsymbol{Q}}_{\rm fea}\left(\boldsymbol{E}^{\rm T}(\boldsymbol{X}\boldsymbol{W}_{\rm v})\right)}_{\text{position contexts}}, \tag{6}$$

where $\tilde{\boldsymbol{Q}}_{\rm fea}$ is the normalized query features $\boldsymbol{Q}_{\rm fea}$ through a standard softmax function along each row (for the same reason in calculating $\tilde{\boldsymbol{K}}_{\rm fea}$). Finally, following [11, 12], we can obtain the output of AugNI:

$$\tilde{\boldsymbol{O}} = \boldsymbol{X} + \tilde{\boldsymbol{Y}}\boldsymbol{W}_{\rm o}. \tag{7}$$

Multi-query AugNI. Inspired by the multi-head [11, 12] mechanism in nonlocal interaction, we intuitively map $\boldsymbol{X}$ into m head subspaces for AugNI. Among each subspace, the model can learn different aspects of feature representations [11, 29]. Following [63], we refer to AugNI under this mechanism as multi-query AugNI. The final object contexts $\tilde{\boldsymbol{O}}_{\rm mn}$ can be obtained through a feature map concatenation of output in each head along the channel dimension:

$$\tilde{\boldsymbol{O}}_{\rm mn} = f_{\rm c}(\tilde{\boldsymbol{O}}_1, \tilde{\boldsymbol{O}}_2, \ldots, \tilde{\boldsymbol{O}}_m)\boldsymbol{W}_{\rm ma}, \tag{8}$$

where $\boldsymbol{W}_{\rm ma} \in \mathbb{R}^{c \times c}$ denotes a learnable linear transformation for multi-query feature fusion.

Compared to the classical nonlocal operation [12], our AugNI contains beneficial relative position information [29, 63]. AugNI also has another advantage of less computational overhead. For example, the

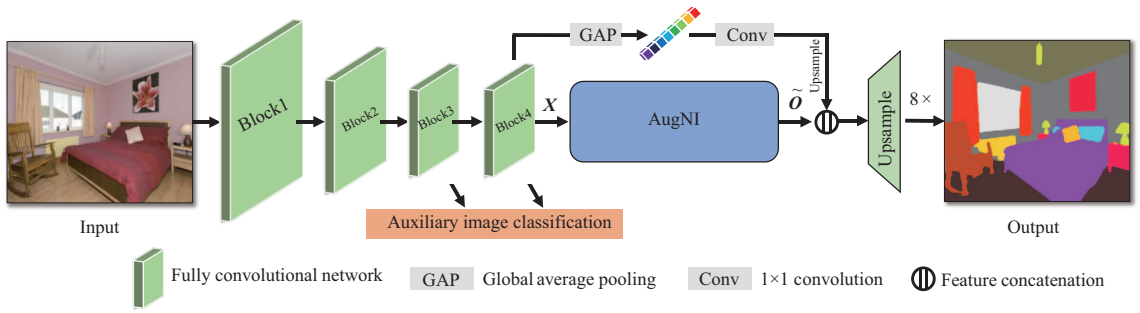


Figure 3 (Color online) An overview of our proposed AugFCN for semantic segmentation. ResNet [38] is used as the backbone, where atrous convolutions are deployed on both block3 and block4 to maintain a high resolution of the output features maps. An image first passes through the atrous ResNet-based FCN; then a set of deep feature maps $\boldsymbol{X}$ can be extracted. After that, our proposed AugNI is implemented for capturing both content- and position-based global object contexts. Besides, two auxiliary branches for multilabel image classification are respectively added into both stage3 and stage4 of the baskbone to improve the model semantics ability. Meanwhile, another auxiliary branch based on a global average pooling layer and a 1×1 convolution layer followed by an upsampling layer is used to capture the global features as in [9]. Finally, the output $\tilde{\boldsymbol{O}}$ of AugNI and the obtained global features are concatenated along the channel dimension for segmentation map predictions after upsampling.

computational complexity of the nonlocal operation is $\mathcal{O}(dh^2w^2)$ because this operation must compute the similarity between each pair of feature grids, while the computational complexity of AugNI is only $\mathcal{O}(mc_{\mathrm{v}}^2)$ because only the similarity between each pair of feature channel is needed. Despite aggregating the global object contexts differently, AugNI and nonlocal operations have a common key property, i.e., the attention intensity summation for each class-wise representation and all other class-wise representations is 1, which represents the normalized feature distribution. Moreover, our AugNI and the channel attention module in DANet [44] mainly differ in the following two points. (1) The motivation is completely different. AugNI is based on a novel interpretation of deep feature maps that aims to capture object contexts, while the channel attention module is based on interdependencies between channel maps and aims to improve the feature representation of specific semantics. (2) AugNI involves the relative position information, unlike the channel attention module.

3.3 AugFCN

On the basis of our proposed AugNI, we further present AugFCN for semantic segmentation in an end-to-end training manner. The overall architecture is illustrated in Figure 3. First, an RGB image is passed through a standard, fully convolutional network (e.g., ResNet [38] and HRNet [68]) as the backbone network. Following [5,23,68], to maintain a high resolution of the output feature maps, atrous convolutions are used on block3 and block4 of the backbone such that the resolution of the output feature maps is 1/8-th of the input image. Then, a group of deep feature maps $\boldsymbol{X}$ can be generated. Next, we apply a 3×3 convolutional layer to reduce the channel dimension of $\boldsymbol{X}$ into 1024. Then, the proposed AugNI is implemented. As in [9,22,23], the channel dimension within AugNI (i.e., c_{k} and c_{v}) is set to 256 for class-wise feature map interactions.

In addition to the main branch described above, following [5,9,16], we add two auxiliary branches for multilabel image classification in stage3 and stage4 of the backbone (i.e., the "auxiliary image classification" branch in Figure 3) to further enrich the feature semantic information. Specifically, the auxiliary classification branch comprises a global average pooling layer for projecting the feature maps into a vector, a 1×1 convolutional layer followed by a ReLU layer for resizing the vector length into 256 and activating this vector, and another 1×1 convolutional layer for resizing the vector length into the class size of the used dataset. Meanwhile, following [9], we implement another auxiliary branch, which is based on a global average pooling layer and a 1×1 convolutional layer followed by an upsampling layer, in stage4 to capture the global image features. Finally, the output feature maps $\tilde{\boldsymbol{O}}$ of AugNI and the obtained global features are concatenated along the channel dimension for semantic segmentation map predictions after being upsampled, where a 3×3 convolutional layer with a batch normalization layer, a ReLU layer, a spatial dropout layer with the drop rate of 0.3, and another 3×3 convolutional layer are added at the end of the concatenated feature maps and before the upsampling layer.

4 Experiments

In what follows, we first introduce two benchmarks and the evaluation metrics (see Subsection 4.1). Then, we show the experimental implementations in detail (see Subsection 4.2). Next, we give the comprehensive ablation studies and result comparisons with state-of-the-art methods on the ADE20K val set (see Subsection 4.3). Finally, we make more result comparisons with state-of-the-art methods on the Cityscapes test set (see Subsection 4.4).

4.1 Benchmarks and evaluation metrics

To demonstrate the superiority of our AugFCN, experiments are performed on two challenging semantic segmentation benchmarks: ADE20K [39] and Cityscapes [20].

• ADE20K [39] is a recently proposed and most challenging semantic segmentation benchmark that contains up to 150 classes of common scenes. This benchmark includes 20k, 2k, and 3k images for the training set, val set, and test set, respectively.

• Cityscapes [20] is a representative, high-resolution (1024×2048) street scene benchmark of 19 classes that has 2975, 500, and 1525 pixel-level annotated images for the training set, val set, and test set, respectively. To make a fair comparison, we only used the finely annotated training images as in [5, 23].

Evaluation metrics. In this work, following the existing segmentation methods [9, 13, 22, 45, 46, 67], we use mIoU and pixel accuracy (Pixel Acc) as the primary metrics. Furthermore, to compare efficiency, the parameters (Params) and GFLOPs are also considered.

4.2 Implementation details

Platform. We implemented all experiments in this paper including ablation studies on EncNet [69] Toolkit[2)] under the PyTorch [70] framework with 8 NVIDIA TITAN Xp GPUs.

Backbone. Following [5, 9, 23], we used classical ResNet [38] with atrous convolutions [10] as our backbone, which was pretrained on the ImageNet [71] dataset, as in most of the previous studies [19,44,69]. ResNet-50 and ResNet-101 were used for ablation studies and result comparisons with state-of-the-art methods, respectively. Moreover, to make a fair comparison with the existing methods on Cityscapes [20], we also used the strong HRNetV2-48 [68] network as the backbone.

Baseline. The standard FCN [4] model with atrous ResNet [10,38] was selected as our base network. Additionally, to make a fair comparison, we also added the global average pooling layer, as in [9], to capture the global image features. We validated the effectiveness of AugFCN on it. Furthermore, we also demonstrated the superiority of our AugNI method over the existing context modeling methods on the baseline FCN.

Comparison methods. To demonstrate its superiority, we compare AugNI with the existing context modeling schemes: PPM [16], ASPP [13], MPM [5], RCC [23], and APNB [21]. Implementing these methods adopts the default settings of their studies. The details are introduced as follows.

• PPM [16]. A four-level, one feature pyramid is first obtained through the global max pooling layer with an output spatial size of 1×1, 2×2, 3×3, and 6×6. Then, a 3×3 convolution is deployed on each layer. Finally, the output features are obtained through feature upsampling and concatenation layers.

• ASPP [13]. ASPP first builds a four-level, one feature pyramid through a 1×1 convolutional layer and three 3×3 atrous convolutional layers [10] under the dilation rate of $[6, 12, 18]$. Then, the feature maps of this feature pyramid are concatenated along the channel dimension as the output feature maps.

• MPM [5]. MPM comprises a lightweight pyramid pooling module (with an output spatial size of 12×12 and 20×20) and a strip pooling module. The feature maps of the lightweight pyramid pooling module are first upsampled into the same spatial size as the input and then concatenated with the feature maps of the strip pooling module. Finally, a 3×3 convolutional layer is added after the fusion operation.

• RCC [23]. A 2-loop version of RCC attention is adopted in our experiments, where three 1×1 convolutional layers are used to generate Q, K, and V feature maps.

• APNB [21]. For APNB, the output spatial size of the pooling layer is set to $[1, 3, 6, 8]$.

Training settings. In the training phase, the batch size was set to 8 for Cityscapes and 16 for ADE20K. Following [5,9,69], the standard cross-entropy loss was used for model optimization, where the loss weight of the auxiliary image classification branch was set to 0.2 and the main segmentation branch

2) https://github.com/zhanghang1989/PyTorch-Encoding.

Table 1 Effectiveness of each component in AugFCN on the val set of ADE20K [39]. "AugNI-c/-p" denotes the content-/position-based AugNI. "AC" denotes the auxiliary multilabel image classification branch. AugFCN is built on ResNet [38] as the backbone by deploying AugNI on the head network and has two auxiliary classification losses.

Setting	AugNI-c	AugNI-p	AC	mIoU (%)	Pixel Acc (%)
Baseline [4]	✗	✗	✗	37.63	77.60
AugFCN	✓	✗	✗	42.30 $_{\uparrow 4.67}$	79.41 $_{\uparrow 1.81}$
AugFCN	✓	✓	✗	43.45 $_{\uparrow 5.82}$	79.87 $_{\uparrow 2.27}$
AugFCN	✓	✓	✓	**43.81**$_{\uparrow 6.18}$	**80.39**$_{\uparrow 2.79}$

was set to 0.8. The "poly" learning rate policy was used with the stochastic gradient descent strategy, where the initial learning rate was multiplied by $1-(\frac{\text{iter}}{\text{max\_iter}})^{\text{power}}$ with power = 0.9. The base learning rate was set to 0.01 for Cityscapes and ADE20K. Following [5, 23], we set the max training epochs to 120 for ADE20K and 180 for Cityscapes. The momentum was set to 0.9, and the weight decay was set to 0.0001. In particular, we also replaced the standard batch normalization with synchronized batch normalization [9] across GPU training, as in [5, 16, 23].

Data augmentation. We followed the same strategy as in [5, 21, 23], and the random scaling was first used on the training set in the range of 0.5–2.0. Then, these training images were randomly cropped to a fixed size of 480×480 for ADE20K and 768×768 for Cityscapes. In addition, random horizontal flips and random brightness jittering were also used.

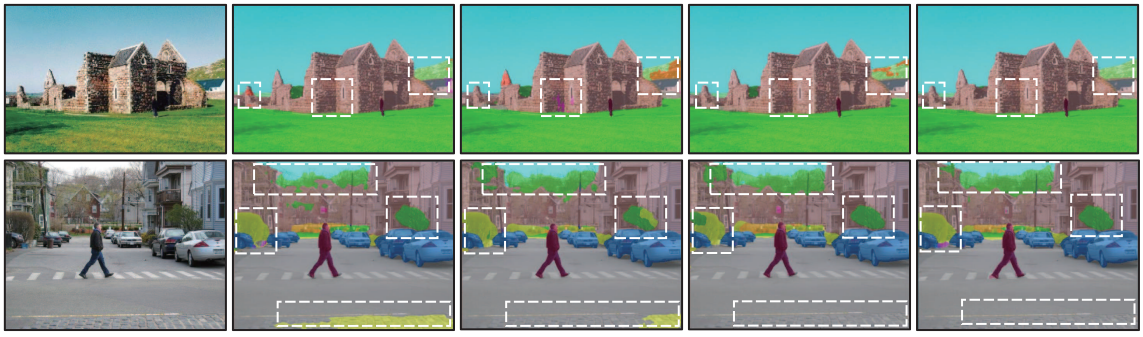
4.3 Experiments on ADE20K

4.3.1 *Ablation studies*

Our ablation studies aim to (a) demonstrate the effectiveness of each component in AugFCN, including position-based AugNI, content-based AugNI, and the auxiliary classification branch; (b) verify whether more position embedding can further improve the model performance; (c) seek out the proper number for multi-query AugNI; (d) make comparisons with the existing context modeling methods; and (e) verify the effectiveness of combining AugNI with other context modeling methods.

(a) Effectiveness of each component in AugFCN. We first decompose the implementation of AugFCN (i.e., AugNI implemented on the classical FCN [4] with two auxiliary multilabel image classification losses) into a content-based component, a position-based component, and an auxiliary image classification component. In Table 1, we list the results for the val set of ADE20K [39] by implementing different components on the baseline model. We observe that each component benefits the model performance. Specifically, implementing the content-based AugNI achieves 42.30% mIoU and 79.41% Pixel Acc, which surpasses the baseline FCN model by 4.67% mIoU and 1.81% Pixel Acc, respectively. This result verifies the effectiveness and importance of the context modeling mechanism in semantic segmentation. Next, we further add our proposed position-based component to the content-based component. The results show that AugNI boosts the model performance by 1.15% mIoU and 0.46% Pixel Acc, which highlights the importance of relative position embedding. In terms of performance, a relative position embedding branch is also conducive to the semantic segmentation model. Finally, with the help of two auxiliary multilabel image classification losses, we can achieve 43.81% mIoU and 80.39% Pixel Acc, which boosts the performance gain by 6.18% mIoU and 2.79% Pixel Acc, respectively, compared to the baseline model.

Qualitative result comparisons of each component with the baseline FCN [4] are visualized in Figure 4. We show segmentation masks on the val set of ADE20K [39] by implementing different components of AugNI on an FCN. We observe that, compared to the baseline, when the content-based AugNI is deployed on an FCN, our AugFCN achieves more accurate segmentation results for some small and distant objects and object parts, e.g., the "`roof`", the "`mountaintop`", and the "`crown`". This finding validates not only the importance of contextual information for semantic segmentation but also the effectiveness of the class-wise feature map interactions in capturing the contextual information. Moreover, when the content- and position-based AugNI are simultaneously deployed on an FCN, our AugFCN works better on some big objects and can generate a more complete mask, e.g., the "`alcazar`", the "`sidewalk`", and the "`tree`". This result confirms the effectiveness of the relative position encoding information in semantic segmentation. When the image classification loss function is added, the imperfection problem in segmentation masks is further reduced, which demonstrates that the classification loss can improve the semantic feature information. AugFCN achieves the best segmentation results when these three components (i.e.,



(a) (b) (c) (d) (e)

Figure 4 (Color online) Qualitative result comparisons with the baseline model [4] on ResNet-50 [38]. Samples are from the val set of ADE20K [39]. "+" denotes implementing the corresponding schemes on the baseline. "AugNI-c/-p" denotes the content-based/position-based AugNI, respectively. "AugNI-all" means implementing all components (i.e., AugNI-c, AugNI-p, and the auxiliary multilabel image classification branch) on the baseline model. The white dashed frames highlight the improved areas predicted by AugFCN. (a) Image; (b) Baseline; (c) +AugNI-c; (d) +AugNI-c & AugNI-p; (e) +AugNI-all.

Table 2 Effectiveness of more position embeddings and the number (m) of multi-query AugNI for AugFCN on the val set of ADE20K [39]. Without an explicit emphasis, the head number is set to 1. "AugNI-2d" denotes implementing the two-dimensional position embedding [29] on AugNI instead of our proposed one.

Setting	AugNI-p	AugNI-2d	mIoU (%)	Pixel Acc (%)
Baseline [4]	✗	✗	37.63	77.60
AugFCN	✓	✗	**43.81** $_{\uparrow 6.18}$	**80.39** $_{\uparrow 2.79}$
AugFCN	✗	✓	43.79 $_{\uparrow 6.16}$	80.32 $_{\uparrow 2.72}$
AugFCN	✓	✓	43.73 $_{\uparrow 6.10}$	80.38 $_{\uparrow 2.78}$
AugFCN ($m = 2$)	✓	✗	43.92 $_{\uparrow 6.29}$	80.60 $_{\uparrow 3.00}$
AugFCN ($m = 4$)	✓	✗	**44.05** $_{\uparrow 6.42}$	**80.78** $_{\uparrow 3.18}$
AugFCN ($m = 8$)	✓	✗	44.11 $_{\uparrow 6.48}$	80.81 $_{\uparrow 3.21}$
AugFCN ($m = 16$)	✓	✗	44.22 $_{\uparrow 6.59}$	80.85 $_{\uparrow 3.26}$

AugNI-c, AugNI-p, and the auxiliary multilabel image classification branch) are deployed simultaneously. The above results confirm the effectiveness of our AugNI.

(b) Will more position embedding be more effective? As stated in Subsection 3.2, one of our key motivations is the lack of relative position information in the existing context modeling methods [23,31], which is not favorable for the highly structured image data. In this subsection, we try to determine whether more relative position embedding will be more effective on semantic segmentation. In particular, the previous, prevalent two-dimensional relative position embedding [29] is used as a comparison method. The experimental results on the val set of ADE20K [39] are shown in the upper part of Table 2. We first compare our proposed position embedding and the two-dimensional position embedding [29] on AugNI, i.e., AugNI-p vs. AugNI-2d. We observe that both relative position embedding methods improve the baseline model performance and almost have the same performance gain, i.e., 43.81% vs. 43.79% on mIoU and 80.39% vs. 80.32% on Pixel Acc. Although AugNI-p slightly outperforms AugNI-2d, we suspect that this victory may be caused by model fluctuation rather than model effectiveness. Furthermore, when AugNI-p and AugNI-2d are simultaneously deployed on AugNI, we observe a lower model performance than previous ones. The reason may be that one relative position embedding method helps the model to obtain enough relative position information, but additional position embedding brings redundant information, which harms semantic segmentation. This result also suggests that we do not need complex structures to design the relative position encoding module and that a learnable relative position embedding module has strong encoding ability. On the basis of this suggestion, in the following experiments, we will use our proposed position embedding in AugNI.

(c) How many queries? On the basis of the base AugNI and inspired by the multi-head in self-attention [11], we also propose multi-query AugNI in Subsection 3.2. In this subsection, we will explore the proper query number (i.e., "m" in Table 2) for multi-query AugNI. We set m to 2, 4, 8, and 16. Additionally, to maintain the channel dimension within AugNI (i.e., c_{k} and c_{v}) as 256, an additional 1×1 convolution is used to adjust the channel dimension if needed. The results on the val set of ADE20K [39] are shown in the lower part of Table 2. Without an explicit emphasis, m is set to 1. We observe that with

Table 3 Effectiveness of AugFCN and result comparisons with the existing context modeling methods on the val set of ADE20K [39]. "+" denotes implementing the corresponding schemes on FCN [4]. Methods separated by dotted-lines belong to different camps.

Setting	mIoU (%)	Pixel Acc (%)
Baseline [4]	37.63	77.60
FCN [4] + PPM [16]	41.50 $_{\uparrow 3.87}$	80.17 $_{\uparrow 2.57}$
FCN [4] + ASPP [13]	42.00 $_{\uparrow 4.37}$	80.23 $_{\uparrow 2.63}$
FCN [4] + MPM [5]	44.03 $_{\uparrow 6.40}$	80.65 $_{\uparrow 3.05}$
FCN [4] + ACF Module [9]	43.67 $_{\uparrow 6.04}$	79.90 $_{\uparrow 2.30}$
FCN [4] + RCC Attention [23]	43.85 $_{\uparrow 6.22}$	80.54 $_{\uparrow 2.94}$
FCN [4] + APNB [21]	43.70 $_{\uparrow 6.07}$	80.38 $_{\uparrow 2.78}$
AugFCN	**44.05** $_{\uparrow 6.42}$	**80.78** $_{\uparrow 3.18}$

increasing m, the performance gain is gradually increased. For example, when m is doubled compared to the standard AugNI, we obtain an extra performance gain of 0.13% mIoU and 0.18% Pixel Acc on AugFCN. When $m = 4$, we obtain 44.05% mIoU and 80.78% Pixel Acc. Particularly, when $m = 16$, we obtain the highest model performance, i.e., 44.22% mIoU and 80.85% Pixel Acc, which surpasses the baseline FCN model by 6.69% mIoU and 3.26% Pixel Acc, respectively. This result suggests that although the divide-and-conquer strategy can help the model learn a stronger representation, too many groupings will not play a big role. Next, considering the model complexity and performance gain, we use a setting of $m = 4$.

(d) Effectiveness of AugNI. To demonstrate the superiority of AugNI in effectiveness, we compare the experimental results of implementing different context modeling methods on an FCN [4]. In this paper, the prevalent context modeling methods used for comparisons are as follows: PPM [16], ASPP [13], MPM [5], ACF module [9], RCC module [23], and APNB [21]. For a fair comparison, we have added an auxiliary, multilabel image classification loss on the backbone, as in [5, 13, 23]. The experimental results on the val set of ADE20K [39] are shown in Table 3. We observe that compared to the baseline, the models in which the context modeling module is added improve the recognition accuracy. Among them, our proposed AugNI has the maximum performance gain of 6.42% mIoU and 3.18% Pixel Acc. Comparatively, PPM has the minimum performance gain of 3.87% mIoU and 2.57% Pixel Acc. The latest MPM achieves performance gains close to AugNI at 6.04% on mIoU and 3.05% on Pixel Acc. In addition to the directly observable conclusion, we also found an interesting phenomenon that the active context modeling methods through long-term dependencies (i.e., MPM, ACF module, RCC module, APNB, and our AugNI) achieve better results on average than the passive methods that aggregate object contexts by increasing the receptive fields (i.e., PPM and ASPP). This phenomenon further confirms the importance of global object contexts in semantic segmentation.

(e) Effectiveness of AugNI and other context modelings. In addition to only adding a context modeling module on an FCN [4], the existing methods (e.g., SPNet [5] and OCNet [24]) empirically confirmed the effectiveness of combining different context modeling methods. Therefore, in this subsection, we explore the possibility of combining AugNI with other context modeling methods, i.e., PPM [16], ASPP [13], MPM [5], APNB [21], ACF module [9], and RCC attention [23]. Specifically, we separately verify the effect of feature map concatenation (i.e., $\oplus$ in Table 4) between AugNI and other context modeling methods and deploy AugNI on each layer of feature maps (i.e., $\odot$ in Table 4) within other context modeling methods. Because the computational process of the ACF and RCC modules does not involve different levels of feature maps, we only do feature concatenation on these two methods. Experimental results on the val set of ADE20K [39] are shown in Table 4. We observe that the help of AugNI, ASPP, the ACF module, and the RCC module can further improve model performance. In particular, deploying AugNI on ASPP boosts the maximum performance gains by 7.30% mIoU and 3.54% Pixel Acc. Conversely, we also observed performance degradation on PPM, MPM, and APNB. A common property of these methods is that they contain pooling operations. Hence, we suspect that the performance degradation on PPM, MPM, and APNB may be due to excessive downsampling operations, which destroy the detailed spatial information.

4.3.2 *Complexity analysis*

In this subsection, we compare model efficiencies by implementing different context modeling methods on the baseline FCN [4], where PPM [16], ASPP [13], and MPM [5] are the representative multi-scale

Table 4 Effectiveness of a joint implementation of our AugNI and other context modeling methods on FCN [4] on the val set of ADE20K [39]. "+" denotes implementing the corresponding schemes on FCN. "⊕" denotes concatenating feature maps of our AugNI and the corresponding context modeling schemes. "⊙" denotes implementing our AugNI on each layer of feature maps within the corresponding context modeling schemes. Methods separated by dotted-lines belong to different camps.

Setting	mIoU (%)	Pixel Acc (%)
Baseline [4]	37.63	77.60
FCN [4] + AugNI (AugFCN)	44.05	80.78
FCN [4] + PPM [16] ⊕ AugNI	41.86 $_{\uparrow 4.23}$	80.32 $_{\uparrow 2.72}$
FCN [4] + PPM [16] ⊙ AugNI	42.07 $_{\uparrow 4.44}$	80.16 $_{\uparrow 2.56}$
FCN [4] + ASPP [13] ⊕ AugNI	44.87 $_{\uparrow 7.24}$	81.00 $_{\uparrow 3.40}$
FCN [4] + ASPP [13] ⊙ AugNI	**44.93** $_{\uparrow 7.30}$	**81.14** $_{\uparrow 3.54}$
FCN [4] + MPM [5] ⊕ AugNI	43.53 $_{\uparrow 5.90}$	80.81 $_{\uparrow 3.21}$
FCN [4] + MPM [5] ⊙ AugNI	43.32 $_{\uparrow 5.69}$	80.73 $_{\uparrow 3.13}$
FCN [4] + APNB [21] ⊕ AugNI	43.72 $_{\uparrow 6.05}$	80.62 $_{\uparrow 3.02}$
FCN [4] + APNB [21] ⊙ AugNI	43.68 $_{\uparrow 6.01}$	80.69 $_{\uparrow 3.09}$
FCN [4] + ACF Module [9] ⊕ AugNI	44.73 $_{\uparrow 6.80}$	80.91 $_{\uparrow 3.31}$
FCN [4] + RCC Attention [23] ⊕ AugNI	44.66 $_{\uparrow 6.83}$	80.90 $_{\uparrow 3.30}$

Table 5 Efficiency analysis on the val set of ADE20K [39]. "+" denotes implementing the corresponding schemes on FCN [4]. Methods separated by dotted-lines belong to different camps.

Setting	Params	GFLOPs
Baseline [4]	27.7 M	77.60
FCN [4] + PPM [16]	↑ 21.0 M	↑ 2.07
FCN [4] + ASPP [13]	↑ 9.8 M	↑ 1.05
FCN [4] + MPM [5]	↑ 11.9 M	↑ 2.02
FCN [4] + ACF Module [9]	↑ 58.1 M	↑ 5.27
FCN [4] + RCC Attention [23]	↑ 10.6 M	↑ 2.38
FCN [4] + APNB [21]	↑ 15.9 M	↑ 4.08
FCN [4] + AugNI	↑ 10.1 M	↑ 2.12

context modeling methods, and APNB [21], the ACF module [9], RCC attention [23], and our AugNI are feature interaction-based methods. The model Params and GFLOPs are used as evaluation metrics. The experimental results on the val set of ADE20K [39] are shown in Table 5. We observe that compared to the baseline FCN [4], models with a context modeling module increase the computational cost and model GFLOPs. Among these methods, the most efficient is ASPP [13], which only brings model Params of 9.8 M and GFLOPs of 1.05. Conversely, the ACF module has the highest computational cost at 58.1 M and GFLOPs at 5.27. The reason is that the ACF module is based on self-attention [11] and nonlocal [12] interaction, which requires a large computational overhead. Moreover, our AugNI has model Params of 10.1 M and GFLOPs of 2.12, which is more efficient than the recent MPM [5], APNB [21], and RCC [23] attention, which are known for their high efficiency in semantic segmentation. In particular, compared to the expensive ACF module [9], AugFCN can reduce up to 48.0 M model parameters and 3.15 GFLOPs, which is extremely significant.

4.3.3 *More experiments*

In addition to applying tricks (e.g., data augmentation and auxiliary training branch) in the training stage, using testing tricks also benefits model performance. In this subsection, we show different experimental settings in the inference stage, including multi-scale (Multi-Scale) and left-right flip (LR-Flip) testing. Experimental results on the val set of ADE20K [39] are shown in Table 6. We observe that Multi-Scale and LR-Flip are conducive to performance improvements. Specifically, using Multi-Scale testing brings a performance gain of 0.58% mIoU and 0.37% Pixel Acc. On the basis of this result, further implementing LR-Flip testing can continuously boost the performance by 0.43% on mIoU and 0.15% on Pixel Acc. When using the stronger ResNet-101 [38] as the backbone, we obtain up to 45.38% on mIoU and 81.65% on Pixel Acc, which surpass the baseline model by 1.33% mIoU and 0.87% Pixel Acc, respectively.

Table 6 More experimental analysis of our AugFCN with different testing tricks and backbones on the val set of ADE20K [39].

Method	Multi-Scale	LR-Flip	mIoU (%)	Pixel Acc (%)
AugFCN-50	✗	✗	44.05	80.78
AugFCN-50	✓	✗	44.63 $_{\uparrow 0.58}$	81.15 $_{\uparrow 0.37}$
AugFCN-50	✓	✓	45.06 $_{\uparrow 1.01}$	81.30 $_{\uparrow 0.52}$
AugFCN-101	✓	✓	**45.38** $_{\uparrow 1.33}$	**81.65** $_{\uparrow 0.87}$

Table 7 Result comparisons with the state-of-the-art methods on the val set of ADE20K [39]. "AugFCN+" denotes deploying AugNI on each layer of feature maps in ASPP [13] module. "–" denotes that there is no reported result in its study. The top three performances are marked as bold, italic, and underline, respectively.

Method	Publication	Backbone	mIoU (%)	Pixel Acc (%)
Baseline (FCN) [4]	CVPR 2015	ResNet-50	34.38	74.57
RefineNet [72]	CVPR 2017	ResNet-101	40.20	–
RefineNet [72]	CVPR 2017	ResNet-152	40.70	–
EncNet [69]	CVPR 2018	ResNet-50	41.11	79.73
GCU [73]	NeurIPS 2018	ResNet-50	42.60	79.51
UperNet [74]	ECCV 2018	ResNet-101	42.66	81.01
CFNet [9]	CVPR 2019	ResNet-50	42.87	–
ACNet [27]	ICCV 2019	ResNet-50	43.01	81.01
PSPNet [16]	CVPR 2017	ResNet-101	43.29	81.39
DSSPN [75]	CVPR 2018	ResNet-101	43.68	81.13
PSANet [76]	ECCV 2018	ResNet-101	43.77	81.51
SAC [77]	ICCV 2017	ResNet-101	44.30	**81.86**
SGR [78]	NeurIPS 2018	ResNet-101	44.32	81.43
EncNet [69]	CVPR 2018	ResNet-101	44.65	*81.69*
GCU [73]	NeurIPS 2018	ResNet-101	44.81	81.19
CFNet [9]	CVPR 2019	ResNet-101	44.89	–
PSPNet [16]	CVPR 2017	ResNet-269	44.94	*81.69*
DANet [44]	CVPR 2019	ResNet-101	45.22	–
CCNet [23]	ICCV 2019	ResNet-101	45.22	–
APNB [21]	ICCV 2019	ResNet-101	45.24	–
APCNet [17]	CVPR 2019	ResNet-101	<u>45.38</u>	–
OCNet [24]	IJCV 2021	ResNet-101	*45.40*	–
AugFCN	None	ResNet-50	45.06	81.30
AugFCN	None	ResNet-101	<u>45.38</u>	<u>81.65</u>
AugFCN+	None	ResNet-101	**45.43**	*81.69*

4.3.4 *Comparisons with the state-of-the-arts*

In this subsection, we compare AugFCN with the state-of-the-art methods[3)] on the val set of ADE20K [39]. Experimental results are given in Table 7 [72–78]. We observe that under ResNet-50, AugFCN achieves a new state-of-the-art performance of 45.06% mIoU and 81.30% Pixel Acc, which surpasses the previous best ACNet [27] by 2.05% on mIoU and 0.29% on Pixel Acc, respectively. When based on a stronger backbone network, ResNet-101, our AugFCN invariably achieves very competitive results of 45.38% mIoU and 81.65% Pixel Acc. Although AugFCN is slightly outperformed by the state-of-the-art OCNet [24] with ResNet-101 (i.e., 45.38% vs. 45.40% mIoU), the object context module of OCNet is intrinsically based on self-attention [11]. Therefore, AugFCN has a plainer computational overhead. Additionally, with the aid of ASPP [13] (i.e., AugFCN+: implementing AugNI on each level of the feature maps of ASPP), we finally obtain a new state-of-the-art performance on the val set of ADE20K of 45.43% mIoU and 81.69% Pixel Acc.

Visualization result comparisons on the val set of ADE20K [39] of AugFCN and the baseline [4] are shown in Figure 5. We show the visualized results by deploying AugNI on different backbone networks, i.e., ResNet-50 (i.e., AugFCN-50), ResNet-101 (i.e., AugFCN-101), and ResNet-101 with ASPP [13] (i.e., AugFCN-101+). Compared to the baseline model, we observe that AugFCN can correct mistakes in object (or some object parts) category predictions, e.g., the "`tree`", the "`sofa`", and the "`desk`".

3) As the advanced transformer-based methods have completely different implementation mechanisms and computational costs than the CNN-based methods, to make a fair comparison, we only make result comparisons with the CNN-based methods.

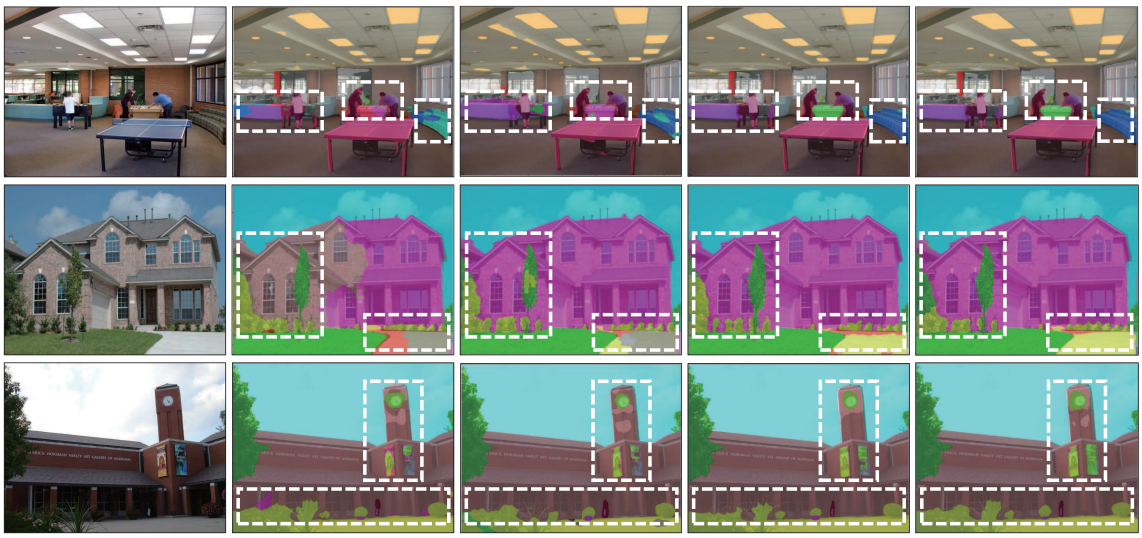


Figure 5 (Color online) Qualitative results on the val set of ADE20K [39]. The baseline model is the standard FCN [4] with atrous ResNet [10]. The white dashed frames highlight the improved areas predicted by our AugFCN. (a) Image; (b) Baseline; (c) AugFCN-50; (d) AugFCN-101; (e) AugFCN-101+.

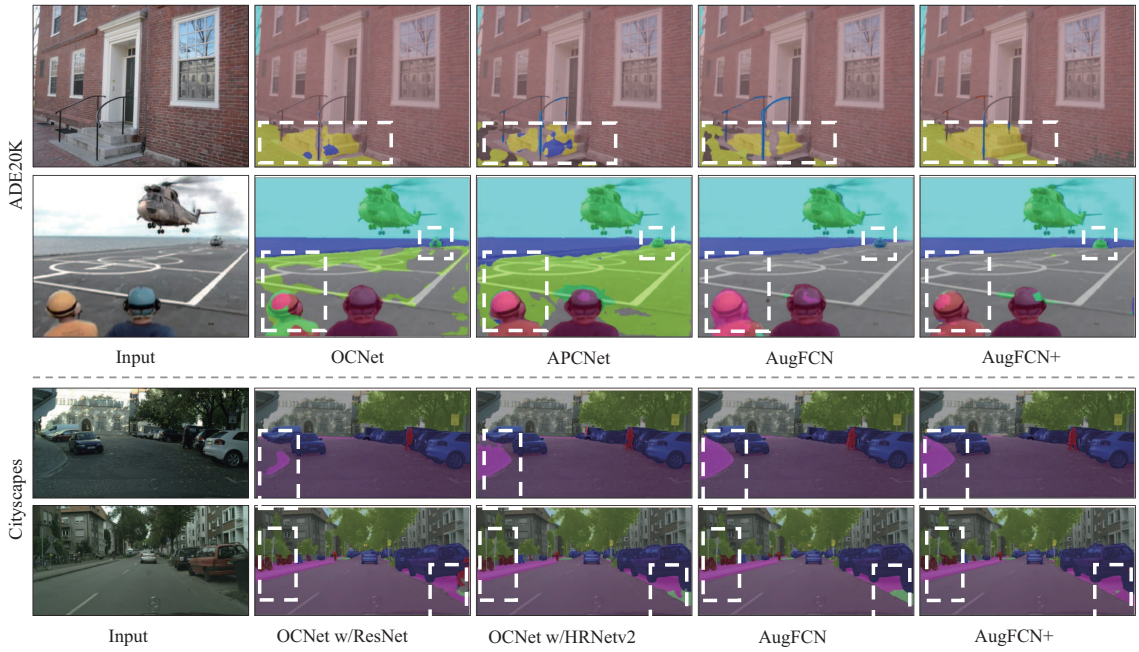


Figure 6 (Color online) Qualitative result comparisons with the state-of-the-art methods on the val sets of ADE20K [39] and Cityscapes [20]. OCNet [24] and APCNet [17] are two state-of-the-art methods on ADE20K dataset, both based on ResNet-101 [38]. OCNet with ResNet-101 (i.e., OCNet w/ ResNet) and with HRNetV2-48 [68] (i.e., OCNet w/ HRNetv2) are two state-of-the-art methods on Cityscapes dataset. The white dashed frames highlight the improved areas predicted by our AugFCN.

These results indicate that effective context information can calibrate the model's false prediction of object category. Furthermore, the large object completeness and small object boundary accuracy can be more accurately predicted, e.g., the "`house`", the "`sidewalk`", and the "`clock`" on the wall. These visualizations empirically verify the effectiveness and importance of contextual information in semantic segmentation tasks. In addition, the last two column visualization results on ResNet-101 and ASPP show that our method can continuously bring further improvements on stronger backbones. In the upper part of Figure 6, we also make qualitative result comparisons between AugFCN and the state-of-the-art OCNet [24] and APCNet [17] on the val set of ADE20K. We observe that AugFCN can predict more accurate segmentation masks in some ambiguous regions, e.g., the "`stair`", the "`head`", and the

Table 8 Result comparisons with the state-of-the-art methods on the test set of Cityscapes [20]. "AugFCN+" denotes that deploying AugNI on each layer of feature maps in ASPP [13] module. The top three performances are marked as bold, italic, and underline, respectively.

Methods	Publication	Backbone	mIoU (%)
DeepLab-v2 [13]	TPAMI 2017	ResNet-101	70.4
DSSPN [75]	CVPR 2018	ResNet-101	77.8
SAC [77]	ICCV 2017	ResNet-101	78.1
DepthSeg [79]	CVPR 2018	ResNet-101	78.2
PSPNet [16]	CVPR 2017	ResNet-101	78.4
ResNet-38 [80]	PR 2019	WiderResNet-38	78.4
PSANet [76]	ECCV 2018	ResNet-101	78.6
BiSegNet [81]	ECCV 2018	ResNet-101	78.9
AAF [82]	ECCV 2018	ResNet-101	79.1
DFN [83]	CVPR 2018	ResNet-101	79.3
CFNet [9]	CVPR 2019	ResNet-101	79.6
PSANet [76]	ECCV 2018	ResNet-101	80.1
DenseASPP [55]	CVPR 2018	DenseNet-161	80.6
SVCNet [84]	CVPR 2019	ResNet-101	81.0
SPGNet [85]	ICCV 2019	2×ResNet-50	81.1
APNB [21]	ICCV 2019	ResNet-101	81.3
BFP [86]	ICCV 2019	ResNet-101	81.4
CCNet [23]	ICCV 2019	ResNet-101	81.4
DANet [44]	ICCV 2019	ResNet-101	81.5
HRNet [68]	TPAMI 2020	HRNetV2-48	81.6
OCNet [24]	IJCV 2021	ResNet-101	81.9
OCNet [24]	IJCV 2021	HRNetV2-48	*82.5*
AugFCN	None	ResNet-50	81.4
AugFCN	None	ResNet-101	81.9
AugFCN+	None	ResNet-101	<u>82.3</u>
AugFCN+	None	HRNetV2-48	**83.0**

"`helicopter`". Because OCNet and APCNet are based on advanced content-based context modeling, we suspect that AugFCN achieves better results than these two methods because it contains more relative position information.

4.4 Experiments on Cityscapes

Quantitative result comparisons with the state-of-the-art methods on the test set of Cityscapes [20] are given in Table 8 [79–86]. We observe that AugFCN with ResNet-50 achieves a new state-of-the-art result of 81.4% mIoU, which greatly surpasses even some previous methods with ResNet-101 as the backbone, e.g., PSANet [76] and CFNet [9].

AugFCN with ResNet-101 can achieve up to 81.9% mIoU, which even surpasses the competitive HRNet [8] with the strong HRNetV2-48 as the backbone by 0.3% mIoU. Under the same backbone network of ResNet-101, AugFCN+ surpasses the state-of-the-art OCNet [24] by 0.4% mIoU (i.e., 82.3% vs. 81.9%). When we use the strong HRNetV2-48 [68] as the backbone, our AugFCN+ achieves a new state-of-the-art performance of 83.0% mIoU, which surpasses OCNet by a large margin of 0.5% mIoU. The above experimental results not only verify the superiority of our AugFCN but also show that our proposed context interaction module has a strong learning ability and can bring consistent performance improvement on a strong backbone network.

Qualitative results on Cityscapes [20] are given in Figure 7. The visualized examples are from the val set of Cityscapes. We show the results on the baseline model, AugFCN-50, AugFCN-101, and AugFCN-101+. Compared to the baseline results, we obtain results similar to those in Subsection 4.3.4., e.g., the "`ground`", the "`sidewalk`", and areas of the "`bonnet`" are better predicted. Furthermore, in the lower part of Figure 6, we also make qualitative result comparisons between AugFCN and the state-of-the-art OCNet [24] on ResNet-101 [38] and HRNetV2-48 on the val set of Cityscapes. We observe that although AugFCN is worse than OCNet in the overall quantitative result, AugFCN achieves better qualitative visualization results on some marginal objects and small objects, e.g., the "`tree`", the "`bicycle`", and

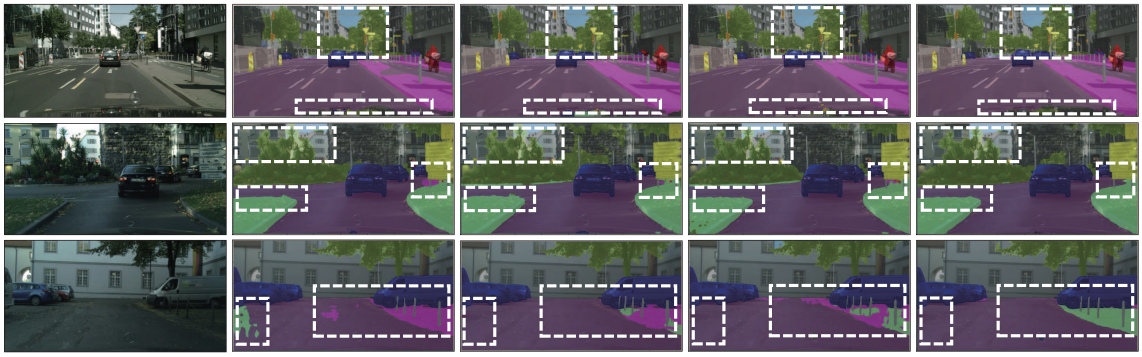


Figure 7 (Color online) Qualitative results on the val set of Cityscapes [20]. The baseline model is the standard FCN [4] with atrous ResNet [10]. The white dashed frames highlight the improved areas predicted by our AugFCN. (a) Image; (b) Baseline; (c) AugFCN-50; (d) AugFCN-101; (e) AugFCN-101+.

the "`trash`". Thus, the effectiveness of our proposed AugFCN is empirically verified.

5 Conclusion and future work

In this paper, we proposed a novel context modeling scheme, termed AugNI, by exploring the content- and position-based object contexts. AugNI stands in the same line as the existing context modeling methods with feature interactions but has less computational overhead. On the basis of AugNI, we further proposed AugFCN for semantic segmentation. The experimental results on two challenging benchmarks, including ADE20K and Cityscapes, show that AugFCN achieves very competitive performance compared to the state-of-the-art context modeling methods. Furthermore, when we combine AugFCN with other context modeling schemes, the experimental results show that AugFCN continuously brings performance improvements to the state-of-the-art context modeling schemes.

AugFCN is based on AugNI, which is designed as the class-wise feature interaction pattern. Inevitably, background information is present in each activated feature map (due to the practical object correlation between foreground objects and the background). Therefore, the class-wise feature interaction is somewhat redundant. In the future, we will continue to study more efficient and specialized context modeling methods, such as aggregating contexts for different categories of objects or using external attributes (e.g., position, shape, and proportion in the image) to establish object contexts. In addition, we will try to apply AugNI to other computer vision tasks, e.g., object detection, person re-identification, and image generation. AugNI is essentially a feature interaction method, so we can also study the application of AugNI to the current popular transformer-based recognition systems to further boost model performance using a vision transformer.

Acknowledgements This work was partially supported by National Key Research and Development Program of China (Grant No. 2018AAA0102002) and National Natural Science Foundation of China (Grant Nos. 61925204, 62172212). The authors would like to thank all the anonymous reviewers for their constructive comments and suggestions.

References

1 Li X, Chen H, Qi X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imag, 2018, 37: 2663–2674
2 Li P, Chen X, Shen S. Stereo R-CNN based 3D object detection for autonomous driving. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019
3 Alhaija H A, Mustikovela S K, Mescheder L, et al. Augmented reality meets computer vision: efficient data generation for urban driving scenes. Int J Comput Vis, 2018, 126: 961–972
4 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
5 Hou Q, Zhang L, Cheng M M, et al. Strip pooling: rethinking spatial pooling for scene parsing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020
6 Feng J P, Wang X G, Liu W Y. Deep graph cut network for weakly-supervised semantic segmentation. Sci China Inf Sci, 2021, 64: 130105

7 Zhang D, Zhang H, Tang J, et al. Self-regulation for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2021

8 Yuan Y, Chen X, Wang J. Object-contextual representations for semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2020

9 Zhang H, Zhang H, Wang C, et al. Co-occurrent features in semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

10 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of International Conference on Learning Representations (ICLR), 2016

11 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2017

12 Wang X, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

13 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 834–848

14 Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2018

15 Ahn J, Cho S, Kwak S. Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

16 Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

17 He J, Deng Z, Zhou L, et al. Adaptive pyramid context network for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

18 Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. 2017. ArXiv:1706.05587

19 Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2018

20 Cordts M, Omran M, Ramos S, et al. The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

21 Zhu Z, Xu M, Bai S, et al. Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019

22 Zhang D, Zhang H, Tang J, et al. Feature pyramid transformer. In: Proceedings of European Conference on Computer Vision (ECCV), 2020

23 Huang Z, Wang X, Huang L, et al. CCNet: criss-cross attention for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019

24 Yuan Y, Wang J. OCNet: object context network for scene parsing. 2018. ArXiv:1809.00916

25 Chen Y, Rohrbach M, Yan Z, et al. Graph-based global reasoning networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

26 Li J, Ma A J, Yuen P C. Semi-supervised region metric learning for person re-identification. Int J Comput Vis, 2018, 126: 855–874

27 Fu J, Liu J, Wang Y, et al. Adaptive context network for scene parsing. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019

28 Ma C, Huang J B, Yang X, et al. Adaptive correlation filters with long-term and short-term memory for object tracking. Int J Comput Vis, 2018, 126: 771–796

29 Bello I, Zoph B, Vaswani A, et al. Attention augmented convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019

30 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of European Conference on Computer Vision (ECCV), 2020

31 Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. In: Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), 2018

32 Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer. In: Proceedings of International Conference on Machine Learning (ICML), 2018

33 Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the coordconv solution. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2018

34 Huang C Z A, Vaswani A, Uszkoreit J, et al. Music transformer. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2018

35 Shen Z, Zhang M, Zhao H, et al. Efficient attention: attention with linear complexities. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2021

36 Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization.

In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017
37 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
38 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
39 Zhou B, Zhao H, Puig X, et al. Scene parsing through ADE20K dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
40 Gao H B, Guo F, Zhu J P, et al. Human motion segmentation based on structure constraint matrix factorization. Sci China Inf Sci, 2022, 65: 119103
41 Zhang Z J, Pang Y W. CGNet: cross-guidance network for semantic segmentation. Sci China Inf Sci, 2020, 63: 120104
42 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations (ICLR), 2014
43 Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
44 Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019
45 Peng C, Zhang X, Yu G, et al. Large kernel matters—improve semantic segmentation by global convolutional network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
46 Zhang Z, Zhang X, Peng C, et al. ExFuse: enhancing feature fusion for semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2018
47 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495
48 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015
49 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015
50 Guo M H, Lu C Z, Liu Z N, et al. Visual attention network. 2022. ArXiv:2202.09741
51 Zhou H, Qi L, Huang H, et al. CANet: co-attention network for RGB-D semantic segmentation. Pattern Recognition, 2022, 124: 108468
52 Zhang D W, Wang B, Wang G R, et al. Onfocus detection: identifying individual-camera eye contact from unconstrained images. Sci China Inf Sci, 2022, 65: 160101
53 Zhang D W, Zeng W, Yao J, et al. Weakly supervised object detection using proposal- and semantic-level relationships. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 3349–3363
54 Zhang D W, Han J, Cheng G, et al. Weakly supervised object localization and detection: a survey. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 5866–5885
55 Yang M, Yu K, Zhang C, et al. DenseASPP for semantic segmentation in street scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
56 Wu T, Tang S, Zhang R, et al. CGNet: a light-weight context guided network for semantic segmentation. IEEE Trans Image Process, 2020, 30: 1169–1179
57 Kong B, Supančič J, Ramanan D, et al. Cross-domain image matching with deep feature maps. Int J Comput Vis, 2019, 127: 1738–1750
58 Li W, Wang X, Xia X, et al. SepViT: separable vision transformer. 2022. ArXiv:2203.15380
59 Chen L, Zhang H, Xiao J, et al. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
60 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
61 Li X, Wang W, Hu X, et al. Selective kernel networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019
62 Zhang H, Wu C, Zhang Z, et al. ResNeSt: split-attention networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022
63 Bello I. LambdaNetworks: modeling long-range interactions without attention. In: Proceedings of International Conference on Learning Representations (ICLR), 2021
64 Tao C, Gao S, Shang M, et al. Get the point of my utterance! Learning towards effective responses with multi-head attention mechanism. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2018
65 Goodfellow I, Bengio Y, Courville A, et al. Deep Learning. Cambridge: MIT Press, 2016
66 Albawi S, Mohammed T A, Al-Zawi S. Understanding of a convolutional neural network. In: Proceedings of International Conference on Engineering and Technology (ICET), 2017
67 Zhong Z, Lin Z Q, Bidart R, et al. Squeeze-and-attention networks for semantic segmentation. In: Proceedings of IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), 2020
68 Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019
69 Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
70 Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2019
71 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009
72 Lin G, Milan A, Shen C, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
73 Li Y, Gupta A. Beyond grids: learning graph representations for visual recognition. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2018
74 Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding. In: Proceedings of European Conference on Computer Vision (ECCV), 2018
75 Liang X, Zhou H, Xing E. Dynamic-structured semantic propagation network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
76 Zhao H, Zhang Y, Liu S, et al. PSANet: point-wise spatial attention network for scene parsing. In: Proceedings of European Conference on Computer Vision (ECCV), 2018
77 Zhang R, Tang S, Zhang Y, et al. Scale-adaptive convolutions for scene parsing. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017
78 Liang X, Hu Z, Zhang H, et al. Symbolic graph reasoning meets convolutions. In: Proceedings of International Conference on Neural Information Processing Systems (NeurIPS), 2018
79 Kong S, Fowlkes C C. Recurrent scene parsing with perspective understanding in the loop. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
80 Wu Z, Shen C, van den Hengel A. Wider or deeper: revisiting the ResNet model for visual recognition. Pattern Recognition, 2019, 90: 119–133
81 Yu C, Wang J, Peng C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2018
82 Ke T W, Hwang J J, Liu Z, et al. Adaptive affinity fields for semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2018
83 Yu C, Wang J, Peng C, et al. Learning a discriminative feature network for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
84 Ding H, Jiang X, Shuai B, et al. Semantic correlation promoted shape-variant context for segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019
85 Cheng B, Chen L C, Wei Y, et al. SPGNet: semantic prediction guidance for scene parsing. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019
86 Ding H, Jiang X, Liu A Q, et al. Boundary-aware feature propagation for scene segmentation. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019