ISNet: Integrate Image-Level and Semantic-Level Context for Semantic Segmentation

Zhenchao Jin, Bin Liu*, Qi Chu, Nenghai Yu
CAS Key Laboratory of Electromagnetic Space Information,
University of Science and Technology of China
{blwx@mail., flowice@, qchu@, ynh@}ustc.edu.cn

Abstract

Co-occurrent visual pattern makes aggregating contextual information a common paradigm to enhance the pixel representation for semantic image segmentation. The existing approaches focus on modeling the context from the perspective of the whole image, i.e., aggregating the image-level contextual information. Despite impressive, these methods weaken the significance of the pixel representations of the same category, i.e., the semantic-level contextual information. To address this, this paper proposes to augment the pixel representations by aggregating the image-level and semantic-level contextual information, respectively. First, an image-level context module is designed to capture the contextual information for each pixel in the whole image. Second, we aggregate the representations of the same category for each pixel where the category regions are learned under the supervision of the ground-truth segmentation. Third, we compute the similarities between each pixel representation and the image-level contextual information, the semantic-level contextual information, respectively. At last, a pixel representation is augmented by weighted aggregating both the image-level contextual information and the semantic-level contextual information with the similarities as the weights. Integrating the image-level and semantic-level context allows this paper to report state-of-the-art accuracy on four benchmarks, i.e., ADE20K, LIP, COCOStuff and Cityscapes [1].

1. Introduction

Semantic image segmentation, which assigns per-pixel predictions of object categories for the given image, is a fundamental problem in computer vision. This task is exceptionally significant to tons of real-world applications, *e.g.*, automatic driving and robot sensing. Recent developments

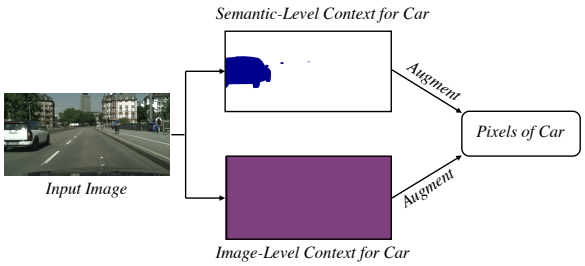


Figure 1. Main idea of integrating image-level and semantic-level context. The blue region on the upper branch denotes for semantic-level context and the purple area on the lower branch stands for image-level context.

of deep neural networks [16, 24] encourage the emergence of a series of works [4, 5, 7, 22, 30, 35], where FCN is the cornerstone of these works. Its encoder-decoder structure, which reduces the spatial dimension to extract features and then leverages upsampling to recover spatial extent, shows numerous improvements in semantic segmentation. Based on this, recent researches mainly focus on two issues to further boost the segmentation performance. One is how to improve the encoder structure so that the models can extract more robust representation for each pixel [25, 29, 34]. The other is how to model the context so that the network can enhance the representation capability of each pixel by encoding the contextual information into the original feature representations [4, 11, 14, 17, 30, 35], which is also the interest of this paper.

Co-occurrent visual pattern inspires the emergence of a series of works about modeling context. These approaches can be roughly divided into two types, *i.e.*, multi-scale context modeling and relational context modeling. For multi-scale context modeling, Deeplab [4] introduces the atrous spatial pyramid pooling (ASPP) so that it can leverage various dilation convolutions to capture the multi-scale contextual information. PSPNet [35] proposes to utilize pyramid spatial pooling to aggregate the multi-scale contextual information. For relational context modeling, Wang *et al.* [26] first revisit the traditional local means [1], and then design

*Corresponding author.
[1]Our code will be available at https://github.com/SegmentationBLWX/sssegmentation.

a non-local block to weighted aggregate the contextual information in the whole image. Zhu *et al.* propose an asymmetric pyramid non-local block to decrease the compution and GPU memory consumption of the standard non-local module. Apart from this, ACFNet [32] and OCRNet [30] first group the pixels into a set of regions, and then augment the pixel representations by weighted aggregating the region representations where the weights are determined by the relations between the pixels and regions. Though impressive, these solutions only focus on aggregating the contextual information from the perspective of the whole image (*i.e.*, image-level contextual information), while discard the significance of the pixel representations of the same category. Accordingly, they all suffer from the same issue that the contextual information of each pixel is unevenly captured from the category region the pixel belongs to and the regions of other categories. For instance, the pixels in the boundaries or the regions of objects of small scales tend to capture much more contextual information from the regions of other objects. Since the label of a pixel is the category of the object that the pixel belongs to, too much contextual information from other objects may cause the network to mislabel these pixels as other categories.

To alleviate the problem above, this paper proposes to augment the pixel representations by aggregating the image-level and semantic-level contextual information, respectively. As illustrated in Figure 1, the image-level context stands for all the pixels in the input image and the semantic-level context denotes for the pixels in the same category region. Based on this definition, an image-level context module (ILCM) is first designed to capture the contextual information from the whole image and thereby, we can obtain the image-level contextual information. Then, a novel semantic-level context module (SLCM) is proposed to aggregate representations of the same category for each pixel (*i.e.*, the semantic-level contextual information), where the category regions are learned under the supervision of the ground-truth segmentation. Next, the similarities between the pixel representation and the image-level contextual information, the semantic-level contextual information are calculated. At last, the pixel representations are augmented by weighted aggregating the image-level contextual information and the semantic-level contextual information, where the weights are determined by the calculated similarities. On the whole, our major contributions are summarized as follows:

- To the best of our knowledge, this paper first explores improving the pixel representations by aggregating the image-level contextual information and semantic-level contextual information, respectively.
- This paper designs a simple yet effective image-level context module (ILCM) and a novel semantic-level context module (SLCM) to capture the contextual information from the perspective of the whole image and the category region, respectively. Experimental results demonstrate the effectiveness of our method.
- A general architecture framework named ISNet is proposed in this paper, which reveals how to leverage ILCM and SLCM to consistently boost the performance of semantic image segmentation. The proposed framework allows this paper to achieve state-of-the-art accuracy on four segmentation benchmarks, *i.e.*, ADE20K, LIP, COCOStuff and Cityscapes.

2. Related Work

Semantic Segmentation. To generate pixel-wise semantic predictions for a given image, image classification networks [4, 24] are extended to yield semantic segmentation masks. FCN [22] is the first work to apply fully convolution on the whole image to produce labels of every pixel and many researchers have made efforts based on FCN in the past few years. Specifically, these studies can be roughly divided into two groups. One is to design a novel backbone network [25, 29] to extract more robust feature representation for each pixel. Considering high-resolution representations are essential for position-sensitive vision problems, Wang *et al.* [25] propose a backbone network named HRNet to maintain high-resolution representations through the whole process. ResNeSt [34] presents a modularized architecture to capture cross-feature interactions and learn diverse representations by utilizing the channel-wise attention on different network branches. The other is to introduce richer contextual information for each pixel [3, 4, 11, 14, 23, 35]. For instance, adopting different sizes of convolutional/pooling kernels or dilation rates to gather multi-scale visual cues [4, 27, 35], employing neural attention [6] to directly exchange the contextual information between paired pixels [3, 11, 17, 28, 31] and building the image pyramids or feature pyramids [18, 21]. This paper focuses on the latter one, *i.e.*, aggregating more meaningful contextual information to augment the pixel representations.

Context Aggregation. While FCN captures information from bottom-up, contextual information with wide field-of-view is also critical for pixel labeling task and is exploited by numerous studies. Deeplab [4, 5, 7] proposes atrous convolution kernels to force the network to perceive larger area and obtain the higher resolution output. Based on Deeplab, DenseASPP [27] densifies the dilated rates to make the network conver larger scale ranges. PSPNet [35] adopts spatial pooling to gain the feature maps of different receptive field sizes so that the combined feature maps could aggregate multi-scale object clues. DANet [11] and OCNet [31] first calculate the similarities between the pixels as the weights, and then improve the pixel representations by weighted ag-

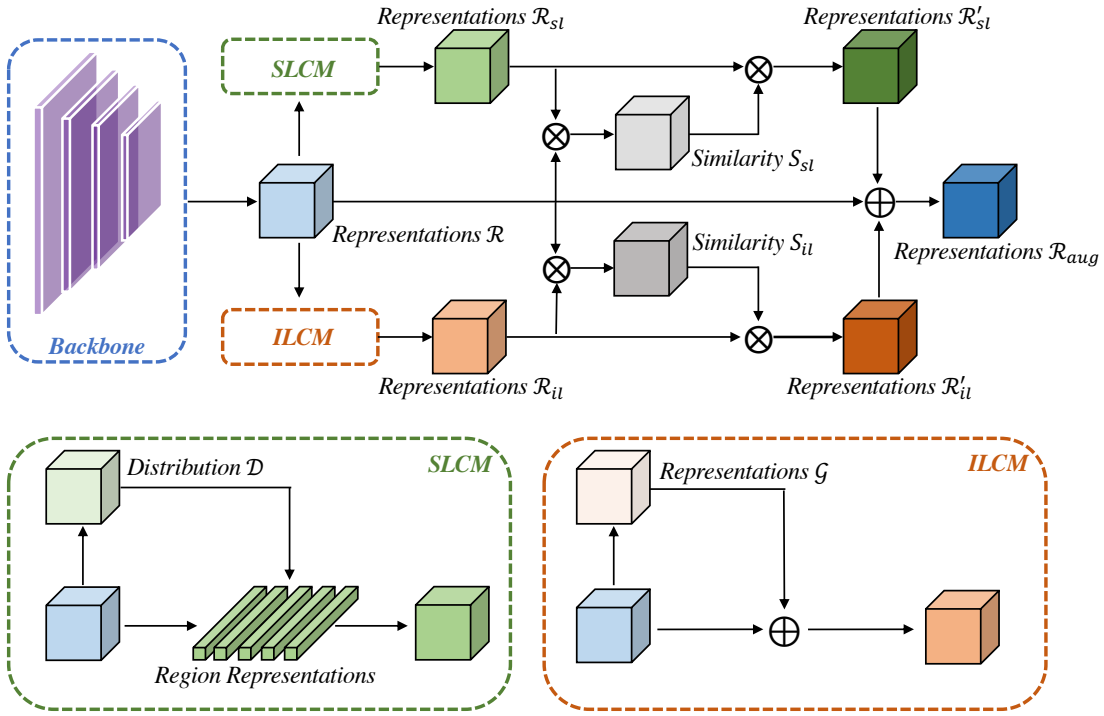


Figure 2. The overview of the proposed framework (ISNet). First, the semantic-level context module (SLCM) and image-level context module (ILCM) are utilized to extract the semantic-level contextual information $\mathcal{R}_{sl}$ and the image-level contextual information $\mathcal{R}_{il}$, respectively. Then, we calculate the similarities between pixel representations $\mathcal{R}$ and $\mathcal{R}_{sl}$, $\mathcal{R}_{il}$. At last, both contextual information are adopted to augment the pixel representations according to the calculated similarities.

gregation of all the pixels in the input image. Apart from this, some works [20, 30, 32] first group the pixels into a set of regions, and then the pixel representations are augmented by weighted aggregating the region representations where the weights are determined by their context relations. Although our semantic-level context module sees similar to these methods, the key difference is that we augment a pixel representation only by adopting the region representation with the same category as the pixel representation rather than all region representations.

3. Methodology

As demonstrated in Figure 2, our ISNet incorporates the image-level contextual information and semantic-level contextual information for semantic image segmentation. We first introduce the overall formulation of our framework in Section 3.1. Then, the details of image-level context module (ILCM) and semantic-level context module (SLCM) are described in Section 3.2 and Section 3.3, respectively. Finally, we show the multi-task loss function used for training the proposed ISNet in Section 3.4.

3.1. Formulation

Given the input image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, we first use a backbone network $\mathscr{B}$ (*e.g.*, ResNet [16]) to project the pixels in $\mathcal{I}$ into a non-linear embedding space so that we can obtain the pixel representations $\mathcal{R}$:

$$\mathcal{R} = \mathscr{B}(\mathcal{I}), \tag{1}$$

where $\mathcal{R}$ is a matrix of size $C \times \frac{H}{8} \times \frac{W}{8}$ and the dimension of a pixel representation is C.

Then, the image-level context module $\mathscr{M}_{il}$ is utilized to aggregate the contextual information from the whole image:

$$\mathcal{R}_{il} = \mathscr{M}_{il}(\mathcal{R}), \tag{2}$$

where $\mathcal{R}_{il}$ is a matrix of size $C \times \frac{H}{8} \times \frac{W}{8}$ storing the image-level contextual information for each pixel representation. Simultaneously, the semantic-level context module $\mathscr{M}_{sl}$ is designed to capture the contextual information within individual category regions:

$$\mathcal{R}_{sl} = \mathscr{M}_{sl}(\mathcal{R}), \tag{3}$$

where $\mathcal{R}_{sl}$ is a matrix of size $C \times \frac{H}{8} \times \frac{W}{8}$ storing the semantic-level contextual information for each pixel representation. After that, we calculate the similarities between $\mathcal{R}$ and $\mathcal{R}_{il}$:

$$\mathcal{S}_{il} = Softmax(\frac{\mathcal{R}^{\frac{HW}{64}\times C} \otimes \mathcal{R}_{il}^{C\times\frac{HW}{64}}}{\sqrt{C}}), \tag{4}$$

where $\mathcal{S}_{il}$ is a matrix of size $\frac{HW}{64} \times \frac{HW}{64}$ and $\otimes$ stands for matrix multiplication. The same for $\mathcal{R}$ and $\mathcal{R}_{sl}$:

$$\mathcal{S}_{sl} = Softmax(\frac{\mathcal{R}^{\frac{HW}{64} \times C} \otimes \mathcal{R}_{sl}^{C \times \frac{HW}{64}}}{\sqrt{C}}), \tag{5}$$

where $\mathcal{S}_{sl}$ is a matrix of size $\frac{HW}{64} \times \frac{HW}{64}$. This operation is inspired by the self-attention mechanism [6]. Next, we leverage $\mathcal{R}_{il}$ and $\mathcal{R}_{sl}$ to augment $\mathcal{R}$:

$$\mathcal{R}_{aug} = \mathscr{A}(\mathcal{R}'_{il} \oplus \mathcal{R}'_{sl} \oplus \mathcal{R}), \tag{6}$$

where $\oplus$ denotes for the concatenation operation and $\mathscr{A}$ is a transform function used to reduce the channels of the input matrix tensors to have size $C \times \frac{H}{8} \times \frac{W}{8}$. $\mathcal{R}'_{il}$ is calculated by using $\mathcal{S}_{il}$ and $\mathcal{R}_{il}$:

$$\mathcal{R}'_{il} = reshape(\mathcal{S}_{il}^{\frac{HW}{64} \times \frac{HW}{64}} \otimes \mathcal{R}_{il}^{\frac{HW}{64} \times C}), \tag{7}$$

and $\mathcal{R}'_{sl}$ is obtained as follows:

$$\mathcal{R}'_{sl} = reshape(\mathcal{S}_{sl}^{\frac{HW}{64} \times \frac{HW}{64}} \otimes \mathcal{R}_{sl}^{\frac{HW}{64} \times C}), \tag{8}$$

where $reshape$ is used to make $\mathcal{R}'_{il}$ and $\mathcal{R}'_{sl}$ have size of $C \times \frac{H}{8} \times \frac{W}{8}$. At last, $\mathcal{R}_{aug}$ is leveraged to predict the labels of the pixels in $\mathcal{I}$:

$$\mathcal{O} = Upsample_{8\times}(\mathscr{H}(\mathcal{R}_{aug})), \tag{9}$$

where $\mathscr{H}$ is a classification head and $\mathcal{O}$ is a matrix of size $K \times H \times W$ storing the predicted class probability distribution of each pixel. K is the number of the categories.

3.2. Image-Level Context Module

Image-level context module $\mathscr{M}_{il}$ is designed to capture the contextual information from the perspective of the whole image. Since there exist co-occurrent visual patterns [35], $\mathscr{M}_{il}$ is widely applied in semantic segmentation task. Prior to this paper, there have been many excellent structures of $\mathscr{M}_{il}$ such as ASPP [4], PPM [35] and OCR [30]. Despite this, since there are two context modules in our framework, we expect the designed $\mathscr{M}_{il}$ in this paper owning the least computation complexity and the increased parameters. Following this expectation, as indicated in Figure 2, we first calculate the channel-wise mean values of the matrix tensor $\mathcal{R}$:

$$\mathcal{G} = \frac{1}{\frac{H}{8} \times \frac{W}{8}} \sum_{ij} \mathcal{R}_{[*,i,j]}, \tag{10}$$

where we leverage the subscript $[i, j]$ or $[*, i, j]$ to index the element or elements of a matrix. $\mathcal{G}$ is a matrix of size $C \times 1 \times 1$ storing the global contextual information of corresponding channels. Then, $\mathcal{G}$ is added into the pixel representations $\mathcal{R}$ to obtain $\mathcal{R}_{il}$:

$$\mathcal{R}_{il} = \mathscr{F}(repeat(\mathcal{G}) \oplus \mathcal{R}), \tag{11}$$

where $\mathscr{F}$ is a transform function used to fuse $\mathcal{G}$ and $\mathcal{R}$, implemented by a 1×1 convolutional layer. $repeat$ is used to repeat the elements in the corresponding channels of $\mathcal{G}$ to make $\mathcal{G}$ have the same shape as $\mathcal{R}$. Note that, the image-level context module can be replaced by all of the existing method such as ASPP [4] and PPM [35] for better modeling the image-level context when pursuing the best segmentation performance. It does not affect the basic motivation of this paper.

3.3. Semantic-Level Context Module

Semantic-level context module $\mathscr{M}_{sl}$ is proposed to aggregate the contextual information within individual category regions. As shown in Figure 2, a classification head $\mathscr{H}'$ is first introduced to predict the category probability distribution $\mathcal{D}$ of the representations in $\mathcal{R}$:

$$\mathcal{D} = \mathscr{H}'(\mathcal{R}), \tag{12}$$

where the size of $\mathcal{D}$ is $K \times \frac{H}{8} \times \frac{W}{8}$ and $\mathscr{H}'$ is implemented by two 1×1 convolutional layers. According to $\mathcal{D}$, the representations in $\mathcal{R}$ can be grouped into various category regions:

$$\mathcal{R}_{c_k} = \{\mathcal{R}_{[*,i,j]} \mid argmax(\mathcal{D}_{[*,i,j]}) = c_k\}, \tag{13}$$

where c_k is between 1 and K standing for the category label and $\mathcal{R}_{c_k}$ is a matrix of size $N_{c_k} \times C$. N_{c_k} denotes for the number of representations belonging to category c_k. For the convenience of presentation, we also define the $\mathcal{D}_{c_k}$ as:

$$\mathcal{D}_{c_k} = \{\mathcal{D}_{[c_k,i,j]} \mid argmax(\mathcal{D}_{[*,i,j]}) = c_k\}, \tag{14}$$

where $\mathcal{D}_{c_k}$ is a matrix of size $N_{c_k} \times 1$. Next, to aggregate the semantic-level contextual information for each pixel representation according to their category, we calculate the region representation for each semantic class c_k as follows:

$$\mathcal{R}'_{c_k} = \sum_{n=1}^{N_{c_k}} \frac{e^{\mathcal{D}_{c_k,[n,*]}}}{\sum e^{\mathcal{D}_{c_k}}} \cdot \mathcal{R}_{c_k,[n,*]} \tag{15}$$

where $\mathcal{R}'_{c_k}$ of size $1 \times C$ is the composite vector of the representations of the same category. After calculating all region representations, we assign them to a matrix tensor according to the class label of corresponding element:

$$\mathcal{R}_{sl,[*,i,j]} = \mathcal{R}'_{c_k} \; if \; argmax(\mathcal{D}_{[*,i,j]}) = c_k \tag{16}$$

where $\mathcal{R}_{sl}$ of size $C \times \frac{H}{8} \times \frac{W}{8}$ is the semantic-level contextual information we demand.

3.4. Loss Function

A multi-task loss function of $\mathcal{D}$ and $\mathcal{O}$ is used to jointly optimize the model parameters. In particular, the loss function of $\mathcal{D}$ is defined as:

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{H \times W} \sum_{i,j} L_{ce}(\mathcal{D}_{[*,i,j]}^{K \times H \times W}, \S(\mathcal{GT}_{[ij]})), \tag{17}$$

where § denotes for converting the ground truth class label stored in $\mathcal{GT}$ into one-hot format and $\mathcal{D}^{K\times H\times W}$ is calculated as follows:

$$\mathcal{D}^{K\times H\times W} = Softmax(Upsample_{8\times}(\mathcal{D})). \tag{18}$$

L_{ce} denotes for the *cross entropy loss* and $\sum_{i,j}$ denotes that the summation is calculated over all locations on the input image I. To let $\mathcal{O}$ contain the accurate category probability distribution of each pixel, we define the loss function of $\mathcal{O}$ as follows:

$$\mathcal{L}_{\mathcal{O}} = \frac{1}{H\times W}\sum_{i,j} L_{ce}(\mathcal{O}_{[*,i,j]},\ \S(\mathcal{GT}_{[ij]})). \tag{19}$$

Finally, we formulate the multi-task loss function $\mathcal{L}$ as:

$$\mathcal{L} = \alpha\mathcal{L}_{\mathcal{D}} + \mathcal{L}_{\mathcal{O}} \tag{20}$$

where α is the hyper-parameters to balance the loss of $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{O}}$. We empirically set $\alpha = 0.4$ by default. With this joint loss function, the model parameters are learned jointly through back propagation.

4. Experiments

4.1. Experimental Setup

Benchmarks. We conduct the experiments on four widely-used semantic segmentation benchmarks.

- **ADE20K** [37] is a scene parsing dataset including 150 categories and diverse scenes with 1,038 image-level labels. This challenging benchmark is divided into 20K/2K/3K images for training, validation and testing.
- **COCOStuff** [2] is a challenging scene parsing dataset that provides rich annotations for 91 thing classes and 91 stuff classes. The dataset contains 9K/1K images for training and testing, respectively.
- **LIP** [13] is a large-scale benchmark that focuses on single human parsing. It contains 50,426 single-person images, which are divided into 30,426 images for training, 10,000 for validation and 10,000 for testing. The pixel-wise annotations cover 19 semantic human part labels and one background label.
- **Cityscapes** [9] is a benchmark for semantic urban scene understanding that contains 19 semantic classes. There are 5K high quality pixel-level finely annotated images and 20K coarsely annotated images in the dataset. The finely annotated 5K images are divided into sets with numbers 2,975, 500, 1,525 for training, validation and testing.

Training Details. We initialize the backbone network using the weights pre-trained on ImageNet and two integrated context modules are initialized randomly. "Poly" learning rate policy with factor $(1 - \frac{iter}{total\_iter})^{0.9}$ is performed for training our framework. Synchronized batch normalization implemented by pytorch is enabled during training. And for the data augmentation, we augment each sample with random scaling in the range of $[0.5, 2]$, random cropping and left-right flipping during training. More specifically, following previous works [30], the training settings for different benchmarks are listed as follows:

- **ADE20K:** The initial learning rate is set as 0.01 and the weight decay is 0.0005. The crop size of the input image is set as 512×512 and batch size is set as 16 by default. The models are fine-tuned for 160K iterations if not specified.
- **COCOStuff:** We set the initial learning rate as 0.001, weight decay as 0.0001, crop size as 512×512, batch size as 16 and training iterations as 60K by default.
- **LIP:** The initial learning rate is set as 0.01 and the weight decay is 0.0005. The crop size of the input image is set as 473×473 and batch size is set as 32 by default. If not specified, the models are fine-tuned for 160K iterations.
- **Cityscapes:** The initial learning rate is set as 0.01 and the weight decay is 0.0005. We set the crop size of the input image as 512×1024, batch size as 8 and the training iterations as 80K if not specified.

Inference Settings. For ADE20K, COCOStuff and LIP, the size of the input image during testing is the same as the size of the input image during training. And for Cityscapes, the input image is zoomed to have its shorter side being 1024 pixels. By default, no tricks (*e.g.*, multi-scale with flipping testing) will be adopted during testing.
Evaluation Metrics. Following the standard setting, mean intersection-over-union (mIoU) is adopted for evaluation.
Reproducibility. The proposed framework is implemented on PyTorch ($version \geq 1.3$) and trained on four NVIDIA Tesla V100 GPUs with a 32 GB memory per-card. And all the testing procedures are performed on a single NVIDIA Tesla V100 GPU. To provide full details of our framework, our code will be made publicly available.

4.2. Ablation Study

ILCM. Since there exist co-occurrent visual patterns, the image-level contextual information is significant for semantic image segmentation. For instance, the car is likely to be in the parking lot or on the highway while not fly in sky. From Table 1, we can see that the image-level contextual module (ILCM) brings an improvement of 5.54% mIoU on

Table 1. Ablation experiments on the image-level context module (ILCM) and the semantic-level context module (SLCM). All methods are learned on the train set of ADE20K and evaluated using single scale test protocal on the validation set.

Baseline	ILCM	SLCM	Backbone	mIoU
✓			ResNet-50	36.96
✓	✓		ResNet-50	42.50
✓		✓	ResNet-50	42.89
✓	✓	✓	ResNet-50	44.09

Table 2. Complexity comparison with existing context schemes. The feature map of size $[1 \times 2048 \times 128 \times 128]$ is adopted to evaluate their complexity during inference. All the numbers are obtained on a single NVIDIA Tesla V100 GPU with CUDA 11.0 and the smaller, the better. As seen, our method requires the least Parameters and the least FLOPs.

Method	Parameters	FLOPs	Time
ASPP [5] (*our impl.*)	42.21M	674.47G	101.44ms
PPM [35] (*our impl.*)	23.07M	309.45G	29.57ms
CCNet [17] (*our impl.*)	23.92M	397.38G	56.90ms
OCRNet [30] (*our impl.*)	14.82M	237.45G	20.22ms
DANet [11] (*our impl.*)	23.92M	392.02G	62.64ms
ANN [38] (*our impl.*)	20.32M	335.24G	49.66ms
DNL [28] (*our impl.*)	24.12M	395.25G	68.62ms
APCNet [15] (*our impl.*)	30.46M	413.12G	54.20ms
ILCM (*ours*)	10.36M	169.77G	42.56ms
SLCM (*ours*)	10.10M	165.47G	53.12ms
ILCM+SLCM (*ours*)	11.02M	180.60G	84.19ms

the validation set of ADE20K. This result demonstrates that ILCM owns the ability of modeling the context from the perspective of the whole image and so that it helps the network better classify the pixels by considering the long-range dependencies.

SLCM. Since the label of a pixel is essentially the category of the object the pixel belongs to, the contextual information from the same semantic class could further enhance the category representation ability of the original pixel representations. Accordingly, the network can leverage the enhanced representations to classify the pixels more accurately. As demonstrated in Table 1, we can see that aggregating the semantic-level contextual information can improve the performance of base framework by 5.93% in terms of mIoU. This improvement well demonstrates the effectiveness of the proposed semantic-level context module.

ILCM+SLCM. Co-occurrent visual pattern makes image-level contextual information be critical for semantic segmentation. However, capturing the contextual information from the whole image will also cause some problems. For instance, since the label of a pixel is decided by the category of the object the pixel belongs to, aggregating too much contextual information from other category regions may cause the network to mislabel the pixel as other categories. To address this issue, this paper proposes to introduce the semantic-level contextual information for each pixel representation additionally, which only leverages the representations in the corresponding category region to enhance each pixel representation. As illustrated in Table 1, we can find that combining ILCM and SLCM outperforms the base model by 7.13% in terms of mIoU. The improvement is much higher than applying single ILCM (7.13% *v.s.* 5.54%) or single SLCM (7.13% *v.s.* 5.93%). This result indicates that ILCM and SLCM can complement and promote each other, which well demonstrates the reliability of the basic motivation, and the effectiveness of the designed framework in this paper.

Table 3. Comparison with the existing context schemes in terms of mIoU. All the models here are learned on the train set of ADE20K and tested on the validation set of ADE20K.

Method	Backbone	Stride	Iterations	mIoU
PPM [35] (*our impl.*)	ResNet-50	8×	160K	42.64
ASPP [5] (*our impl.*)	ResNet-50	8×	160K	43.19
OCR [30] (*our impl.*)	ResNet-50	8×	160K	42.47
ANN [38] (*our impl.*)	ResNet-50	8×	160K	41.75
NonLocal [26] (*our impl.*)	ResNet-50	8×	160K	42.15
DNL [28] (*our impl.*)	ResNet-50	8×	160K	43.50
CCNet [17] (*our impl.*)	ResNet-50	8×	160K	42.47
ISNet (*ours*)	ResNet-50	8×	160K	**44.09**

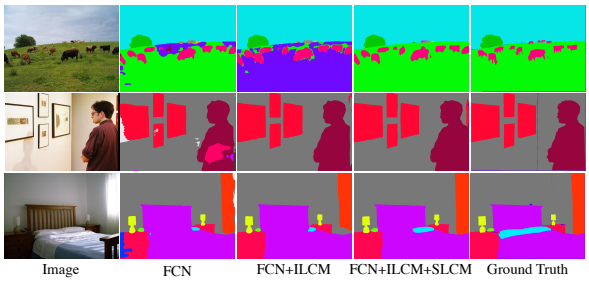


Figure 3. Qualitative results on the validation set of ADE20K. All the models here are trained under the same setting with ResNet-50 as the backbone network. Best viewed in color and zoom in.

Qualitative Results. Figure 3 shows some qualitative results to further prove the reliability of our basic motivation. As seen, the segmentation performance has been well improved after introducing the semantic-level context module (SLCM). For instance, in the 2nd row, the painting with only a small visible part on the right side of the image is still well segmented (the ground truth is wrong). This result well indicates that aggregating too much contextual information from other category regions may cause the network to mislabel one pixel as other categories, and introducing SLCM can well alleviate this problem.

Complexity. Table 2 demonstrates the complexity comparison with existing context schemes, including the increased parameters, computation complexity (measured by the number of FLOPs) and inference time. As seen, the proposed context scheme requires the least parameters and the least computation complexity. Specifically, ILCM+SLCM only requires $\frac{1}{4}$ and $\frac{1}{2}$ of the parameters of ASPP and PPM respectively, which can prevent our model from overfitting to a certain extent. Furthermore, ILCM+SLCM only re-

Table 4. Segmentation results on ADE20K validation set. Multi-scale and flipping testing is employed here for fair comparison. The best score is marked in **bold**.

Method	Backbone	Stride	mIoU
PSPNet [35]	ResNet-101	8×	43.29
PSANet [36]	ResNet-101	8×	43.77
EncNet [33]	ResNet-101	8×	44.65
OCNet [31]	ResNet-101	8×	45.08
OCRNet [30]	ResNet-101	8×	45.28
CCNet [17]	ResNet-101	8×	45.76
ANNet [38]	ResNet-101	8×	45.24
ACNet [12]	ResNet-101	8×	45.90
DMNet [14]	ResNet-101	8×	45.50
APCNet [15]	ResNet-101	8×	45.38
DANet [11]	ResNet-101	8×	45.22
OCRNet [30]	HRNetV2-W48	4×	45.66
ISNet (*ours*)	ResNet-50	8×	45.04
ISNet (*ours*)	ResNet-101	8×	47.31
ISNet (*ours*)	ResNeSt-101	8×	**47.55**

Table 5. Comparison of performance on the test set of COCOStuff with state-of-the-art approaches. Multi-scale and flipping testing is leveraged here for fair comparison.

Method	Backbone	Stride	mIoU
OCRNet [30]	ResNet-101	8×	39.50
SVCNet [10]	ResNet-101	8×	39.60
DANet [11]	ResNet-101	8×	39.70
EMANet [20]	ResNet-101	8×	39.90
SpyGR [19]	ResNet-101	8×	39.90
ACNet [12]	ResNet-101	8×	40.10
OCRNet [30]	HRNetV2-W48	4×	40.50
ISNet (*ours*)	ResNet-50	8×	40.16
ISNet (*ours*)	ResNet-101	8×	41.60
ISNet (*ours*)	ResNeSt-101	8×	**42.08**

quires $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{7}{10}$, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{2}{5}$ of the FLOPs based on PPM, ASPP, CCNet, OCRNet, DANet, ANNet, DNL and APCNet respectively. These results well prove the efficiency of the proposed method.

Performance Comparison. To further show the necessity of introducing the semantic-level contextual information, we compare the performance of ISNet with the existing context schemes under the same training and testing settings. As illustrated in Table 3, we can see that ISNet outperforms all the existing context modules by yielding a mIoU of 44.09%. Noted that, the image-level context module integrated into ISNet only gains a mIoU of 42.50% which is much weaker than most of the existing image-level context schemes. This result well demonstrates the effectiveness of introducing the semantic-level contextual information.

4.3. Comparison with State-of-the-Art

ADE20K. Results of other state-of-the-art semantic segmentation solutions on ADE20K are summarized in Table 4. As is known, ADE20K is challenging due to its various image scales, plenty of semantic classes and the gap between its training and validation set. Even under such circumstance, ISNet employing ResNet-50 achieves a mIoU of 45.04%, which is 1.75%, 1.27% and 0.39% mIoU higher than PSPNet [35], PSANet [36] and EncNet [33] using a stronger ResNet-101 backbone network, respectively. This result further shows the importance of aggregating the semantic-level contextual information for augmenting each pixel representation. Furthermore, as we can see, the previous best method named ACNet achieves a mIoU of 45.90%. Our ISNet with ResNet-101 achieves superior mIoU of 47.31% which is 1.41% mIoU higher than the previous state-of-the-art. Besides, integrating ILCM and SLCM also allows this paper to report new state-of-the-art performance on the validation set of ADE20K, *i.e.*, 47.55% by leveraging ResNeSt-101.

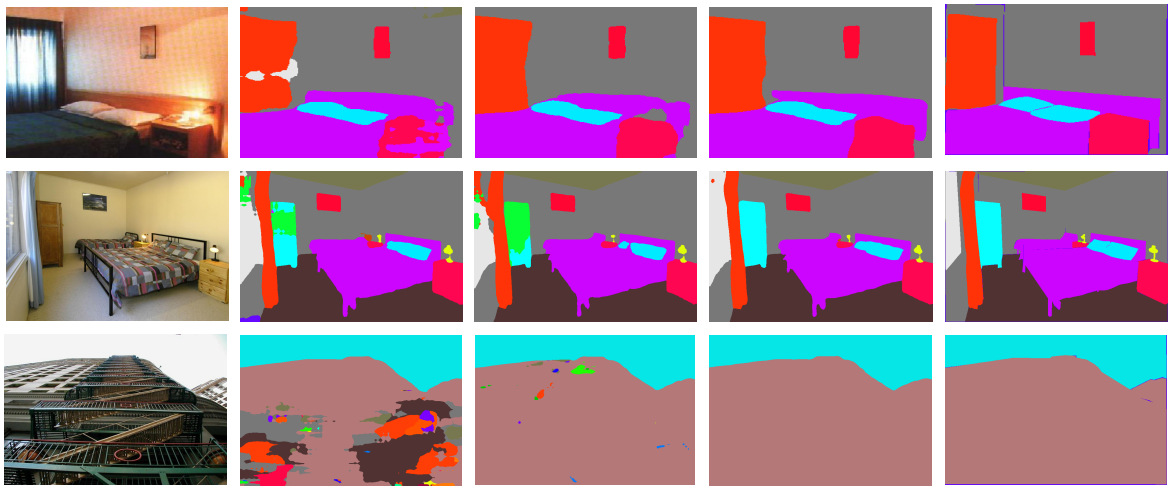
Table 6. State-of-the-art comparison on the validation set of LIP. Flipping testing is utilized here for fair comparison. ‡ means that we adopt ASPP as the image-level context module.

Method	Backbone	Stride	mIoU
DeepLab [4]	ResNet-101	-	44.80
CE2P [23]	ResNet-101	16×	53.10
OCRNet [30]	ResNet-101	8×	55.60
OCNet [31]	ResNet-101	8×	54.72
CCNet [17]	ResNet-101	8×	55.47
HRNet [25]	HRNetV2-W48	4×	55.90
OCRNet [30]	HRNetV2-W48	4×	56.65
ISNet (*ours*)	ResNet-50	8×	53.41
ISNet (*ours*)	ResNet-101	8×	55.41
ISNet (*ours*)	ResNeSt-101	8×	56.81
ASPP (*our impl.*)	ResNet-101	8×	55.34
ISNet‡ (*ours*)	ResNet-101	8×	**56.96**

Table 7. Segmentation results on Cityscapes validation set. Only single-scale testing is adopted here.

Method	Backbone	Stride	mIoU
GCNet [3, 8]	ResNet-101	8×	79.03
PSPNet [8, 35]	ResNet-101	8×	79.76
PSANet [8, 36]	ResNet-101	8×	79.31
ANN [8, 38]	ResNet-101	8×	77.14
NonLocal [8, 26]	ResNet-101	8×	78.93
CCNet [8, 17]	ResNet-101	8×	78.87
EncNet [8, 33]	ResNet-101	8×	78.55
DANet [8, 11]	ResNet-101	8×	80.41
DNL [8, 28]	ResNet-101	8×	80.41
OCRNet [8, 30]	HRNetV2-W48	4×	80.70
ISNet (*ours*)	ResNet-50	8×	79.32
ISNet (*ours*)	ResNet-101	8×	80.56
ISNet‡ (*ours*)	ResNet-101	8×	**81.10**

COCOStuff. Since there are only 9K images in the train set of COCOStuff where these images contain 182 semantic classes, COCOStuff is a very challenging benchmark for scene parsing. Table 5 compares the performance of the state-of-the-art methods. By leveraging ResNet-50 as the backbone network, ISNet achieves a mIoU of 40.16%, which is already higher than the most of the previous state-of-the-art methods. When leveraging the same backbone



(a) Image (b) FCN (c) PSPNet (d) ISNet (e) Ground Truth

Figure 4. Qualitative results on the validation set of ADE20K. All the models here are trained under the same setting. Best viewed in color and zoom in.

network ResNet-101, our ISNet surpasses OCRNet [30], EMANet [20], DANet [11] and ACNet [12] in a large margin, *i.e.*, 2.10%, 1.70%, 1.90% and 1.50% mIoU, respectively. Furthermore, because of the effectiveness of integrating image-level and semantic-level context for semantic image segmentation, our ISNet with ResNeSt-101 reports the new state-of-the-art performance on the testing set of COCOStuff, *i.e.*, 42.08%.

LIP. LIP is a fine-grained semantic segmentation benchmark which has additional challenges such as the complex clothes texture, the scale diversity of different categories, the deformable human body, and the fine-grained segmentation of labels. Therefore, it is hard to model the image-level context by only leveraging the proposed ILCM, which simply concatenates output features of an average pooling layer. Despite this, as demonstrated in Table 6, ISNet with ResNet-101 still achieves a mIoU of 55.41% which is very competitive among the previous state-of-the-art methods. To report a new state-of-the-art performance, we replace the original ILCM with ASPP [4] here to better model the image-level context. As seen, our ISNet‡ outperforms the previous best method OCRNet with HRNetV2-W48 by 0.31% mIoU while ASPP only achieves 55.34% in terms of mIoU which is 1.62% lower than our ISNet‡. These results consistently prove the basic motivation of this paper, *i.e.*, additionally introducing the semantic-level contextual information is critical to improving the pixel representations.

Cityscapes. As shown in Table 7, we also show the comparative results with other state-of-the-art methods on the validation set of Cityscapes. We can see that our model ISNet‡ with ResNet-101 is superior to previous best method OCRNet with HRNetV2-W48. Specifically, integrating the image-level context module ASPP and the proposed semantic-level context module SLCM makes this paper report a new state-of-the-art with mIoU hitting 81.10% under single-scale testing. This result further proves the rationality of aggregating the image-level and semantic-level context for augmenting each pixel representation.

Qualitative Results. Figure 4 illustrates the qualitative results on the validation set of ADE20K. We can see that our ISNet could achieve better segmentation results than both FCN (*i.e.*, without context module) and PSPNet (*i.e.*, using the image-level context module), which further shows the effectiveness of our method (*i.e.*, adopting both image-level and semantic-level context module).

5. Conclusion

This paper studies the context aggregation problem. Motivated by the fact that the existing image-level context schemes may bring too much contextual information of other categories into the pixel representations so that it makes the network mislabel the pixel, we propose to integrate the image-level contextual information and semantic-level contextual information, respectively, to further boost the performance of semantic segmentation. Specifically, we first design a simple yet effective image-level context module as a common practice to capture the global semantic structured information. Then, the semantic-level contextual information is also aggregated for each pixel by leveraging the proposed semantic-level context module. At last, the pixel representations are augmented by weighted aggregating the image-level contextual information and the semantic-level contextual information. Extensive experiments demonstrate the effectiveness of our method. Integrating image-level and semantic-level context allows us to report new-state-of-arts on four segmentation benchmarks, *i.e.*, ADE20K, Cityscapes, LIP and COCOStuff.

References

[1] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[8] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[10] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[12] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6757, 2019.

[13] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[14] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3562–3572, 2019.

[15] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[18] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.

[19] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8950–8959, 2020.

[20] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019.

[21] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[23] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4814–4821, 2019.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution represen-

tation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[27] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[28] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

[29] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.

[30] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

[31] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[32] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6798–6807, 2019.

[33] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[34] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[36] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018.

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[38] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019.