Interaction via Bi-directional Graph of Semantic Region Affinity for Scene Parsing

Henghui Ding[1,3] † Hui Zhang[1] Jun Liu[2] Jiaxin Li[3] Zijian Feng[3] Xudong Jiang[1]
[1]Nanyang Technological University [2]Singapore University of Technology and Design [3]ByteDance

Abstract

In this work, we devote to address the challenging problem of scene parsing. It is well known that pixels in an image are highly correlated with each other, especially those from the same semantic region, while treating pixels independently fails to take advantage of such correlations. In this work, we treat each respective region in an image as a whole, and capture the structure topology as well as the affinity among different regions. To this end, we first divide the entire feature maps to different regions and extract respective global features from them. Next, we construct a directed graph whose nodes are regional features, and the bi-directional edges connecting every two nodes are the affinities between the regional features they represent. After that, we transfer the affinity-aware nodes in the directed graph back to corresponding regions of the image, which helps to model the region dependencies and mitigate unrealistic results. In addition, to further boost the correlation among pixels, we propose a region-level loss that evaluates all pixels in a region as a whole and motivates the network to learn the exclusive regional feature per class. With the proposed approach, we achieves new state-of-the-art segmentation results on PASCAL-Context, ADE20K, and COCO-Stuff consistently.

1. Introduction

Scene parsing (or semantic segmentation), as one of the most fundamental tasks in computer vision, targets at segmenting an image to different regions and assigning each region a specific class label. Parsing errors exist widely in previous methods, due to the diverse appearances and the complicated topology structures among objects. In this paper, we propose an approach that constructs affinity dependency among different semantic regions, helps to reason global affinities among objects/stuff in the given image and mitigates the deviated segmentation results.

The scene parsing can be regarded as a pixel-level classification task (*i.e.*, recognition) as well as a region cluster task (*i.e.*, segmentation). Previous works pay more attention to recognition than region cluster, resulting in "spot" predictions, ambiguous results and unrealistic results, as shown in Figure 1. Previous attempts to these issues mainly revolved around capturing large receptive fields for each pixel, like the pyramid [8, 81] and the non-local [20] receptive fields. These context methods, though build implicit connections among different pixels, aim at aggregating context for each pixel and thereby do not take advantage of the region-level correlations.

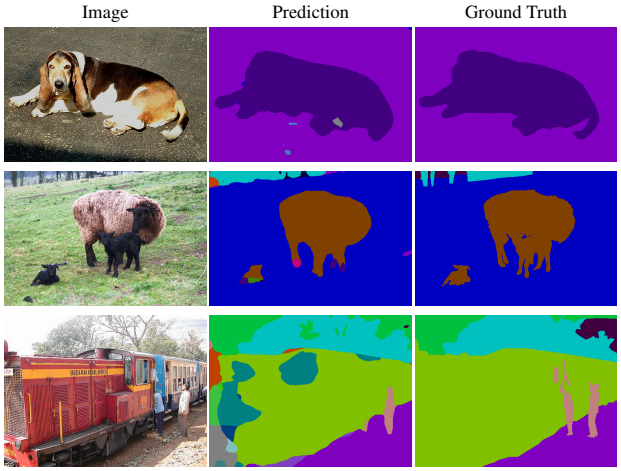

Figure 1: Common errors in scene parsing: "spot", ambiguous, and unrealistic predictions.

In this work, we focus more on region clustering and attribute the errors in Figure 1 to the lack of region-level constrains for each pixel and insufficient regional correlations. Concretely we boost scene parsing by applying region-level constraints and establishing connections among regions. We split the feature maps to various regions and regard pixels in each respective region as a whole to explore the region-level correlations. This approach helps cluster the features from the same region and thus remove the "spot" pieces in prediction. To this end, we first generate a coarse segmentation mask (*e.g.*, the

†Henghui Ding (ding0093@ntu.edu.sg) is the corresponding author.

prediction in Figure. 1) to define the regions in feature maps. We then propose a graph-based region affinity reasoning (GRAr) module to place region-level constraints onto these split regions. The occurrence of the error spots and ambiguous/unrealistic results is significantly reduced.

The edges are vital for determination of the connections among nodes in a graph. Unlike previous graph-based works [10, 40, 27] that learn edges from scratch, we utilize affinity information from training samples via statistics. Different affinities among semantic classes are observed, as some objects frequently co-occur in images, while some objects never appear together. Therefore, it is advantageous to model the complicated affinities and spatial dependencies among different objects. However, the class-affinity has not been well investigated in previous graph methods. Here, we calculate the coexistence times between each two classes and list them as a confusion matrix to represent the affinities. For each category, its supporters and opponents are achieved by examining the confusion matrix, intuitively watching who supports and who suppresses its existence. For instance, the class "pillow" in ADE20K [83], has supporters of "bed" and "sofa" and opponent of "bus". Using these category affinity information, we capture the topology structure among different regions and the affinity dependency of different categories. A directed graph is constructed then, in which each node represents a semantic region and each edge represents the directed affinity connection between the two nodes.

Furthermore, we propose a semantic region loss to facilitate the region-level feature clustering. From the discussion in [4], the training objectives of FCN-based segmentation networks is always based on the assumption that pixels are independent. Whereas, it is also well known that each pixel in a given scene image is highly correlated with other pixels, and treating them independently during training fails to utilize the correlation among pixels. Some context works exploit such correlation implicitly, while their training objective functions still regard pixels as independent. In this work, we propose a semantic region loss (SR-Loss) that treat pixels in the same region as a whole to explicitly boost the inner correlations. The SR-Loss formulates a region-level recognition task and prompts the network to learn the regional-level features for respective class.

The main contributions of this paper are summarized as follows:

- We propose a bi-directional graph according to statistics of class-correlations in training samples, and infer region affinity based on this graph.
- We provide the computed affinity-aware features for corresponding regions to improve feature representations and mitigate the unrealistic results.
- We propose a semantic region loss that offers region-level recognition supervision, motivating the network to learn the discriminative region-level features for each class.
- The proposed approach achieves new state-of-the-art performance consistently on three popular scene parsing benchmarks, PASCAL-Context, ADE20K, and COCO-Stuff.

2. Related Work

2.1. Scene Parsing

Scene parsing, or semantic segmentation, is one of the challenging and fundamental tasks in computer vision. Recently, deep-learning-based scene parsing methods have achieved excellent progress, benefiting from the great success of deep convolutional neural networks on computer vision [37, 25, 62, 46, 65, 63, 50, 13, 45, 18, 66, 77, 30]. The seminal work FCN [49] introduces the fully convolutional networks (FCN) to semantic segmentation. Plenty of FCN-based segmentation works are proposed then, including the encoder-decoder approaches (*e.g.*, DecovNet [52], U-Net [54], EFCN [57], CGBNet [17] and SegNet [3]) that extract high-level features by encoder and then gradually restore the spatial details by decoder, and Dilated-FCN [8, 71, 81] that retains more spatial details in encoder, by discarding some downsampling operations in CNNs and utilizing dilated convolutions to compensate the receptive fields.

Contextual modeling plays a vital role in scene parsing. A plenty of works in segmentation focus on aggregating better context. Multi-scale pyramid representation is one of the common methods. For example, the DeepLab [8] proposes an Atrous Spatial Pyramid Pooling module, known as ASPP; the PSPNet [81] introduces Pyramid Pooling Module (PPM) to capture multi-scale contextual information from different regions; and the DenseASPP [68] utilizes denser dilated rates to cover much larger scale ranges. Self-attention [61, 64], as another popular method, has representatives like the DANet [20] which applies non-local operation over both spatial and channel dimensions, while CCNet [28] enables non-local attention lightweight by decomposing the non-local attention into two consecutive criss-cross attentions.

Unlike the previous methods that aggregate context for each pixel, our method focuses on regional-level affinity and treats each region in a given image as a whole. In addition, we captured the structural topology and affinity between different sub-regions of the input image to enhance well-parsed regions and correct wrongly predicted pieces. Besides, our approach shares the spirit of coarse-to-fine strategies, by employing coarse segmentation mask to divide the image to different regions and construct

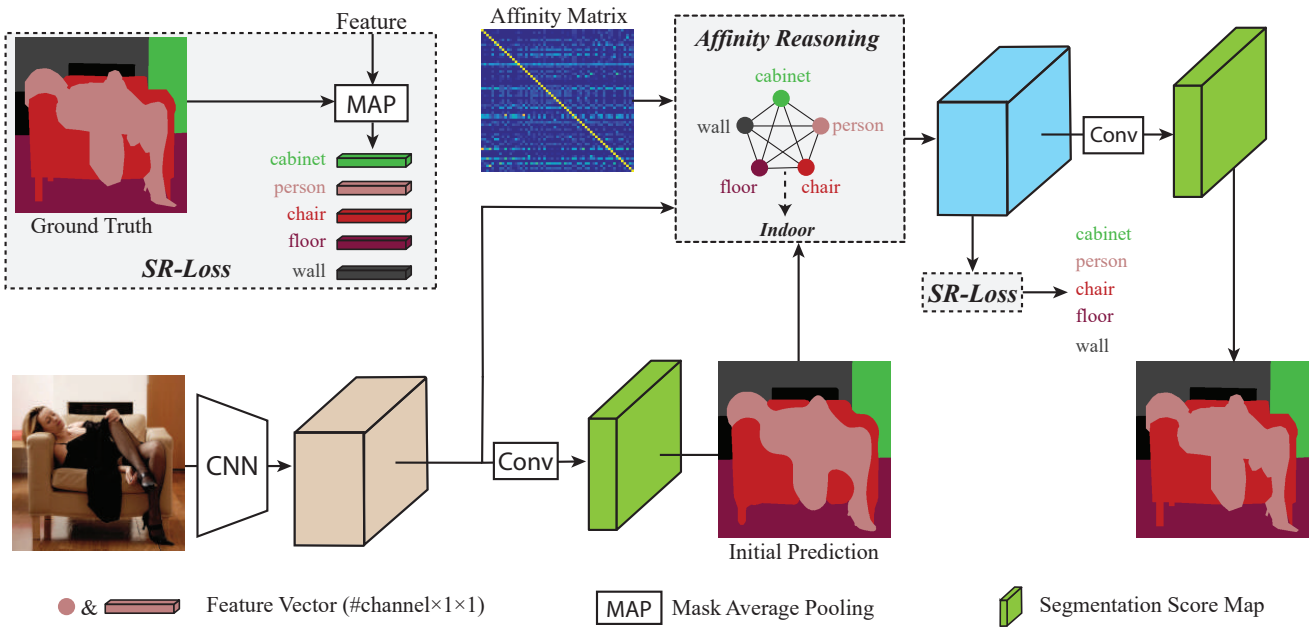

Figure 2: The overall architecture of our proposed approach. We first employ a segmentation network to generate the initial segmentation mask, and use this prediction to split the feature maps to different regions. Next, we feed these region features to our graph-based region affinity reasoning (GRAr) module to boost the region representation of these features under the guidance of an affinity matrix, which is collected via doing statistics on datasets. Further, Semantic Region Loss (SR-Loss), which extracts class-specific features using ground truth mask, is employed as region-level classification supervision to boost the correlations of pixels within the same semantic region.

graph based on these regions, as differed from previous LRN [33], AFNet [74] and OCR [72] that utilize coarse-to-fine strategy to refine their initial segmentation prediction. Our approach is flexible in the input division approach of input image, with other methods like super-pixel [1, 31] and contour detection [2, 7, 5] also compatible.

2.2. Graphical Models

Graphic models have a long history in computer vision and have drawn lots of attentions till now. Probabilistic graphical model [22, 39, 38, 36, 8, 48], *e.g.*, Markov Random Fields (MRF) and Conditional Random Fields (CRF), is proposed to capture the long-range dependencies based on undirected graphs. DAG-RNN [58], SPN [47] and BFP [14] employs directed acyclic graphs to recurrently propagate information. AAF [34] devotes to replacing the pixel-wise cross-entropy loss by adversarial learning loss to capture affinity based on CRF. Graph Convolution Networks (GCN) [35, 12] is proposed to conduct convolution on graphs and have have achieved great success in computer vision tasks [11]. Several GCN-based methods are proposed for semantic segmentation [10, 40, 27]. Different from previous scene parsing methods using graphical models, in this work, we utilize affinity knowledge to establish a bi-directional graph for the affinity reasoning among different semantic regions.

3. Approach

The main innovation of this work is that we treat every region as a whole, and capture the structure topology and affinity among different regions of the input image to enhance the well-parsed pieces and correct the wrongly-parsed pieces. We first propose a directed graph whose nodes are features of semantic regions and edges are affinities between these regions, in the section 3.1 and section 3.2. Specifically, we explain how to derive the affinities among regions from the given directed graph, in section 3.2. Then, we define a semantic region loss in section 3.3 to facilitate the network to learn discriminative features for each semantic region. The overall architecture is shown in Figure 2. We employ a segmentation network to generate the coarse segmentation mask, which is used to extract node representations from corresponding regions. The proposed directed graph is utilized for affinity reasoning. Semantic region loss and segmentation loss are both adopt for supervision.

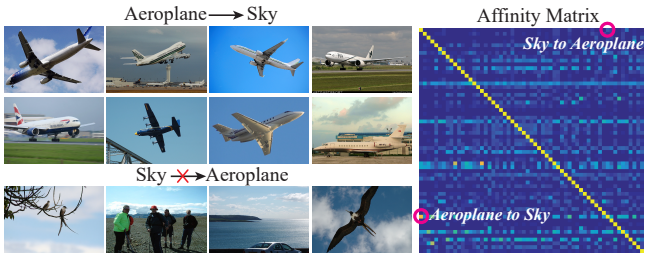

Figure 3: The connections are directed. The aeroplane strongly supports the existence of sky, but the sky only weakly support the existence of aeroplane.

3.1. Semantic Region Inferring

To begin with, we explain our approach of semantic region inferring. The split regions are extracted respectively to be the nodes in graphs. Previous graph-based methods, *e.g.*, GloRe [10] and SpyG [40], do not define explicit nodes for their graph reasoning. Instead, their graph reasoning simply leaves the problem to the model itself and relies on implicit nodes. It is rather difficult to capture what kind of graph the model has learned and how the information flows in feature maps. Each pixel may act as a node in the learned graph, under which case, the graph reasoning works in the same way as non-local attention [20, 64]. As opposed to previous graph reasoning, we apply region-level constrains and treat each region as a separate node in our graph. By this way, we give a clear definition for the graph nodes.

There are alternative ways to segment the feature maps and define the regions, *e.g.*, superpixel. We adopt segmentation mask because it is easily obtained from our network itself. Once the roughly-predicted mask is obtained, we utilize it to define the regions, and use the proposed affinity reasoning to refine the prediction in reverse. It is inherently a coarse-to-fine progress, in which a coarse prediction is obtained before the final refined prediction. The detailed process is as follows. We divide the feature maps into respective parts according to the initial predicted segmentation mask. Average global pooling is then performed on each split region to extract region-level features, which constitute the node representations for our directed graph. The region affinity is inferred based on these region-level features under mutual restraint, as explained below.

3.2. Semantic Region Affinity Reasoning with Bi-directional Graph

Here, we illustrate how to construct the connecting edges in the bi-directional graph based on the acquired node representations, and then conduct the Graph-based Region Affinity Reasoning (GRAr). We observe that in existing segmentation datasets like PASCAL-Context [51] and ADE20K [83], the pre-defined categories can be clustered to different groups, *e.g.*, indoors and outdoors. Moreover, some categories are frequently co-occurrent, *e.g.*, cow and grass, while some categories never appear together, *e.g.*, bed and airplane. Hence, we consider collecting category affinity in the segmentation data set by statistics, and use this affinity to construct directed edges in the graph. We count the co-occurrence times of every two classes, and calculate their co-occurrence frequency by:

$$f_{i,j} = \begin{cases} \dfrac{t_{i,j}}{\sum_k^N t_{k,j} - t_{j,j}}, & i \neq j, \\ 1, & i = j, \end{cases} \tag{1}$$

where $t_{i,j}$ is co-occurrence times of class i and class j, and $f_{i,j}$ represents the frequency that class i is co-occurrent with class j. The co-occurrence frequency matrix of Pascal-Context [51] is shown in Figure 3. If category i frequently co-occurs with class j, we assume it as one of the supporters of class j and assign stronger connections to them in the graph. These connections are directed because such supports are not bilateral. For example, the aeroplane strongly supports the existence of the sky, because the appearance of the aeroplane is always accompanied by the sky as a background, while the sky only supports the existence of the aeroplane weakly, because the sky also co-occurs frequently with other objects like bird and people, as shown in Figure 3. Therefore, for every two categories, *e.g.*, i and j, there are bi-directional connections, $f_{i,j}$ and $f_{j,i}$, that represent the support probability from i to j and from j to i respectively.

We use the affinity matrix to construct the directed edges in graph. From the affinity matrix, we infer two different affinity edges, one is positive and another is negative. The positive edge represents the support from the frequent co-occurrent objects, *e.g.*, aeroplane and sky, while the negative one represents the suppress from the objects that hardly co-occurs, *e.g.*, grass and bed. The positive edge from i to j is $e^p_{i,j}$ and $e^p_{i,j} = f_{i,j}$, the corresponding negative edge is $e^n_{i,j}$ and $e^n_{i,j} = 1 - f_{i,j}$. We perform affinity reasoning using the edges in the graph (see Figure 4). Suppose there are N nodes in our graph, and the representation for each node is F_j, where $j \in \{1, 2, ..., N\}$. The positive affinity reasoning for each node is derived by:

$$F^p_j = F_j + \lambda Conv(\sum_{i=1}^{N} F_i e^p_{i,j}, \Theta) \tag{2}$$

where $e^p_{i,j}$ represents the positive affinity edge from i^{th} node to j^{th} node, *e.g.*, from airplane to sky. $Conv$ is a convolution and Θ is its parameter, we add a residual skip and λ is set to its learnable parameter. The existence of a node's class is supported by receiving positive affinities

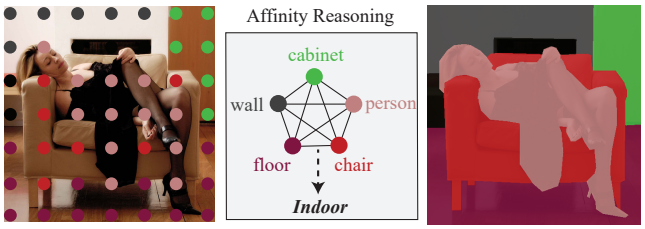

Figure 4: We extract region features according to coarse segmentation mask, and these region features are used as node representations in our graph. The graph edges are bi-directional affinities between each two nodes. We conduct affinity reasoning based on this directed graph and map the nodes features back to their corresponding regions.

from other nodes, which indicates that when the class is mapped back, the corresponding regional features are enhanced.

In addition to the positive affinity, we also introduce a negative affinity that inhibits the existence of a class. For example, the "airplane" and "grass" indicate outdoor scene and their co-occurrence suppresses the existence of indoor objects like "bed". Specifically, the negative affinity is used to remove some unrealistic pieces/spots in the coarse prediction. The negative affinity reasoning for each node can be formulated as:

$$F_j^n = F_j - \hat{\lambda} Conv(\sum_{i=1}^{N} F_i e_{i,j}^n, \hat{\Theta}) \tag{3}$$

where $e_{i,j}^n$ represents the negative affinity edge from i^{th} node to j^{th} node. The feature expression of mapping a node to a specific category will be somehow weakened, if it obtains negative affinity from other nodes, which helps correct the wrongly predicted pieces/spots in the coarse prediction. The node representation is concatenated with its positive and negative affinity reasoning to perform the final representation,

$$\dot{F}_j = Conv(F_j \oplus F_j^n \oplus F_j^p, \dot{\Theta}) \tag{4}$$

where $\oplus$ denotes concatenation. Finally, we map these nodes features $\dot{F}_j$ back to their corresponding regions. By affinity reasoning, the features in the same region are better clustered, and also capture the context from other regions. Moreover, the correctly predicted pieces are boosted while incorrectly predicted pieces are suppressed and rectified.

3.3. Semantic Region Loss

It is more widely accepted that every pixel in a given scene image is highly correlated with other pixels, despite the claim that the training of FCN-based segmentation networks is based on the assumption of pixel independence. Therefore, treating pixels independently in training results wastes pixel correlations. Previous context-related works [8, 81, 20] exploit pixel correlations implicitly, while keeping objective training functions pixel-independent. We, however, propose a semantic region loss (SR-Loss) that regard pixels in the same region as a whole, for boosting the pixel correlations. The proposed semantic region loss motivates the network to learn discriminative region-level features for each class, and meanwhile clusters features that belongs to the same class.

To extract global features for each class that presents in the given image, we perform global mask average pooling over CNN features, using the ground truth segmentation mask:

$$\hat{\mathbb{F}}_k = \frac{1}{\sum 1(\mathbb{M} = k)} \sum 1(\mathbb{M} = k)\mathbb{F} \tag{5}$$

where $\mathbb{M}$ is ground truth segmentation mask, $1(*)$ is the binary indicator that outputs 1 when $*$ is $True$. $\mathbb{F}$ is CNN features with spatial size of $H \times W$, $\hat{\mathbb{F}}_k$ is a vector with spatial size of 1×1 that represents global features for class k. Then we feed $\hat{\mathbb{F}}_k$ to an additional fully connected layer with softmax for classification of class k, supervised by the cross-entropy loss $\mathcal{L}_k$. If there are K classes that exist in the given image, there will be $\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_K$, *i.e.*, each class that presents in the given image is assigned with a region-level classification loss. Each SR-Loss $\mathcal{L}_k$ acts on a certain region that corresponds to class k.

As differed from pixel-level cross-entropy loss in FCN segmentation methods, our proposed Semantic Region Loss treats the pixels with same label as a whole, which facilitates the network to learn the class-specific clustering features and enlarges the feature distinction among various categories. It is a region-level classification between the pixel-level and the image-level. Our SR-loss also differs from the SE-Loss in EncNet [75] which treats all categories as a multi-label image-level classification problem, in that we treat each semantic region as a single-label multi-class image classification. Suppose these are K classes in the given image, the final loss is:

$$\mathcal{L} = \mathcal{L}_s + \frac{w_{sr}}{K} \sum_{k=1}^{K} w_k \mathcal{L}_k \tag{6}$$

where $\mathcal{L}_s$ is pixel-wise segmentation loss, w_{sr} is the weight for our SR-Loss, K is the number of classes existing in the given image, w_k is the loss weight for $\mathcal{L}_k$.

4. Experiments

4.1. Implementation Details

All our experiments are based on the open source platform Pytorch [53]. We test our approach based on

ResNet [25] and ResNest [78] (pre-trained on ImageNet [55]). The CNN backbone is truncated from last Pooling layer and the last two downsampling operations are discarded, *i.e.*, the output stride is 8. Dilated convolutions are used to compensate the shrinkage of receptive fields caused by discarding the down-sampling. The network is fully convolutional and trained end-to-end with mini-batch at a batchsize of 16. Batch Normalization [32] is used after new added convolutional layers to accelerate training. For training data augmentation, we flip the images horizontally, resize the images between 0.5 and 2, and rotate them between -10 and 10 degrees, randomly. Following prior works [8, 81, 70, 76], we adopt the "poly" scheduling to adjust the learning rate: $lr_c = lr_b \times (1 - \frac{iter}{total\_iter})^{0.9}$, where lr_c is the current learning rate and lr_b is the initial base learning rate, $total\_iter$ is the number of total training iterations and $iter$ is current iteration step. Momentum and weight decay are fixed to 0.9 and 0.0001 respectively.

4.2. Datasets and Evaluation Metrics

We report our results on 3 challenging scene parsing benchmarks, PASCAL-Context, ADE20K, and COCO-Stuff. All three of them provide pixel-wise segmentation mask with diverse scenes and categories. We collect their affinity matrix by performing statistics over all images.

- **PASCAL-Context** [51] provides dense pixel-wise segmentation maps for whole scenes. It has 10103 scene images from Pascal VOC [19]. There are 4998 training images and 5105 testing images. The most frequent 59 object/stuff categories and background are used for evaluation. In training, we resize and crop images from PASCAL-Context as 544×544 for batch processing. Batch size is set to 16, the base learning rate lr_b is set to 0.001 and $total\_iter$ is 50K.

- **ADE20K** [83] contains 20210 training images, 2000 validation images and 3352 test images. In this dataset, there are 150 categories, including 35 stuff categories and 115 discrete objects categories, annotated to each pixel. In training, images are resized and cropped to 544×544 for batch processing. Batch size is set to 16, the base learning rate lr_b is set to 0.01 and $total\_iter$ is 200K.

- **COCO-Stuff** [6] provides detailed pixel-level annotations for 10000 images from Microsoft COCO dataset [44]. There are 9000 training images and 1000 testing images. In Microsoft COCO segmentation dataset [44], images are annotated with 80 objects labels, the unlabeled stuff concepts are further labeled with the new added 91 stuff categories in COCO-Stuff. We report our results on 171 categories, including all the objects and stuff categories. In our training, images are resized and cropped to 544×544 for batch processing. Batch size is set to 16, the base learning rate lr_b is set to 0.001 and $total\_iter$ is 100K.

Backbone	GRAr	SR-Loss	MS	mIoU%
ResNet-50				42.3
ResNet-50	✓			51.7
ResNet-50	✓	✓		53.0
ResNet-101	✓	✓		54.8
ResNet-101	✓	✓	✓	55.7
ResNest-101	✓	✓	✓	57.0

Table 1: Ablation Study on PASCAL-Context. Baseline is dilated FCN, MS means multi-scale testing.

Backbone	Pos	Neg	Affinity Method	mIoU %
ResNet-50			$N.A.$	42.3
ResNet-50	✓		$Conv$	48.1
ResNet-50		✓	$Conv$	47.9
ResNet-50	✓	✓	$Conv$	48.2
ResNet-50	✓		$Graph$	49.7
ResNet-50		✓	$Graph$	49.3
ResNet-50	✓	✓	$Graph$	51.7

Table 2: Ablation Study of the proposed Affinity Reasoning with Directed Graph.

Method	Parameters	Memory	Time	mIoU%
PPM [81]	23.2M	226M	75ms	49.2
ASPP [8]	15.5M	81M	74ms	49.9
DANet [20]	10.6M	668M	101ms	50.6
OCR [72]	10.5M	93M	41ms	50.4
GRAr (ours)	2.4M	95M	42ms	51.7

Table 3: FCN+"module" comparison. We compare with the plug-in "module" of some previous methods in terms of efficiency/effectiveness based on our re-implementations. The size of input feature map is $1 \times 2048 \times 68 \times 68$.

We evaluate the proposed segmentation network with mean Intersection-over-Union (**mIoU**), please refer to [49] for its mathematical definition.

4.3. Ablation Study

We conduct ablation studies to showcase the effectiveness of each module employed in the proposed approach. First, as shown in Table 1, compared to the baseline FCN, our bi-directional Graph-based Region Affinity Reasoning (GRAr) module brings a mIoU performance improvement of 9.4% on PASCAL-Context, which affirms the capability of affinity inference to aggregate global clues and further enhance segmentation results. Next, using SR-Loss, we

achieve a mIoU gain of 1.3% in performance, which shows that our segmentation network learn more precise class-wise representations under the region-level classification supervision. The proposed affinity reasoning extract useful clues from coarse regions and correct the wrongly predicted pieces. Additionally, we test our approach based on different backbone, and demonstrate that stronger backbone can extract more representative features and enhance the finial segmentation performance.

In order to further investigate the proposed affinity reasoning module in detail, we conduct another ablation study in Table 2. We study the affinity method, *i.e*. these two positive and negative affinity in Eq (2) and Eq (3). Firstly We use a simple $Conv$, *i.e*., discard the $e^p_{i,j}$ in Eq (2), in this case our affinity reasoning is totally learnable and initial from scratch. As shown in Table 2, $Graph$ is 1.6% better than $Conv$ based on positive affinity only, and is 1.4% better than $Conv$ based on both positive and negative affinity reasoning. The results show that our graph edges provide more explicit affinities among different nodes, while the $Conv$ that learns from scratch without any supervision only captures limited and fuzzy affinities. Next, we conduct experiments to study the positive and negative affinities. Positive affinity is employed to boost the correctly predicted regions, while negative affinity is used to suppress the incorrectly predicted pieces. They work together to enhance the final segmentation prediction. As shown in Table 2, single graph, *i.e*., positive-only or negative-only, already reaches encouraging results, for example, positive-affinity-only achieves mIoU of 49.7%. When we adopt positive affinity and negative affinity at the same time, further performance improvement is obtained, which shows that positive affinity and negative affinity benefit each other. It is easy to understand by taking an example that, when there is one "spot" with wrong label, negative affinity removes the high response of the wrong-class's features and positive affinity improve the feature response of correct class with clues from other regions.

To have a fair comparison with previous works, we re-implement some state-of-the-art works based on our backbone with ResNet-50. We conduct a FCN + "module" comparison in Table 3. Besides the channel adaptation layer and the final classifier, our approach mainly introduces three new 1×1 convolution layers in Eq. (2) to (4) and is indeed light-weight. We also report the increased parameters, GPU memory, inference time by adding these modules to backbone. The comparison in Table 3 demonstrates the superiority of our proposed approach in terms of both efficiency and effectiveness.

4.4. Comparison with State-of-the-Art Works

In this section, we present our segmentation results on bencharmks and comparision with state-of-the-art works.

Methods	Backbone	mIoU %
FCN-8s [56]	VGG16	39.1
PixelNet [4]	VGG16	41.4
DAG-RNN [58]	VGG16	43.7
FCRN [67]	VGG16	44.5
DeepLab-v2[8]	ResNet101	45.7
Global-Context[29]	ResNet101	46.5
RefineNet [43]	ResNet101	47.1
PSPNet [81]	ResNet101	47.8
CCL [15]	ResNet101	51.6
EncNet [75]	ResNet101	51.7
DUpsampling [60]	Xception-71	52.5
DANet [20]	ResNet101	52.6
SpyGR [40]	ResNet101	52.8
EMANet [42]	ResNet101	53.1
BFP [14]	ResNet101	53.6
CPNet [69]	ResNet101	53.9
HRNet [59]	HRNetV2-W48	54.0
ACNet [21]	ResNet101	54.1
SPNet [26]	ResNet101	54.5
RecoNet [9]	ResNet101	54.8
OCR [72]	ResNet101	54.8
OCR [72]	HRNetV2-W48	56.2
Ours	ResNet101	**55.7**
Ours	ResNest101	**57.0**

Table 4: Testing results on **PASCAL-Context**.

The proposed approach achieves new state-of-the-art segmentation results consistently on COCO-Stuff, ADE20K and PASCAL-Context.

PASCAL-Context. We test our segmentation network over 59 categories, our results and previous state-of-that-art works are shown in Table 4. It can be seen that the proposed approach outperforms previous methods based on ResNet-101. And we further test our approach based on stronger backbone ResNest-101, this achieves the best mIoU performance 57.0%, outperforms the 56.2% of OCR based on HRNetV2-W48.

ADE20K testing results are presented in Table 5. We report our results on 2000 validation images. It can be seen that the proposed approach achieves new state-of-the-art performance, 47.1% based on ResNet-101 and 47.9% based on ResNest-101, outperforming existing methods.

COCO-Stuff testing results are shown in Table 6. We test the proposed scene parsing approach over 171 catefories, and report our results on 1000 validation images. As shown in Table 6, the proposed approach outperforms previous methods. We achieve 41.9% based on ResNet-101 and 42.6% based on ResNest-101, significantly outperforming previous state-of-the-art OCR [72] based on HRNetV2-W48

Qualitative Results. We illustrate some segmentation examples from PASCAL-Context, ADE20K, and COCO-Stuff, as shown in Figure 5. The second column and

Networks	Backbone	mIoU %
SegNet[3]	VGG16	21.6
FCN [56]	VGG16	29.4
DilatedNet [71]	VGG16	32.3
DAG-RNN [58]	VGG16	33.5
RefineNet [43]	ResNet152	40.7
PSPNet [81]	ResNet101	42.0
PSANet [82]	ResNet101	43.8
SAC [80]	ResNet101	44.3
EncNet [75]	ResNet101	44.7
SFNet [41]	ResNet101	44.7
CFNet [79]	ResNet101	44.9
CCNet [28]	ResNet101	45.2
ANNet [84]	ResNet101	45.2
APCNet [24]	ResNet101	45.4
OCNet [73]	ResNet101	45.5
DMNet [23]	ResNet101	45.5
RecoNet [9]	ResNet101	45.5
SPNet [26]	ResNet101	45.6
OCR [72]	HRNetV2-W48	45.7
CPNet [69]	ResNet101	46.3
Ours	ResNet101	**47.1**
Ours	ResNest101	**47.9**

Table 5: Testing results on **ADE20K**.

Networks	Backbone	mIoU %
FCN [6]	VGG16	22.7
DeepLab [8]	VGG16	26.9
DAG-RNN [58]	VGG16	30.4
RefineNet [43]	ResNet101	33.6
CCL [15]	ResNet101	35.7
OCR [72]	ResNet101	39.5
SVCNet [16]	ResNet101	39.6
DANet [20]	ResNet101	39.7
EMANet [42]	ResNet101	39.9
SpyGR [40]	ResNet101	39.9
ACNet [21]	ResNet101	40.1
OCR [72]	HRNetV2-W48	40.5
Ours	ResNet101	**41.9**
Ours	ResNest101	**42.6**

Table 6: Testing results on **COCO-Stuff**.

third column are our baseline's results and the proposed approach's prediction. It can be seen that the proposed approach sigificiantly imporve the segmentation peformance of our baseline, which shows that our approach could correct most of the wrongly predicted pieces in predictions of baseline.

5. Conclusion

In this work, we address the problem of scene parsing from a new perspective. We regard all the pixels in the same semantic region as a whole, based on the recognition that pixels in a given image, especially pixels belonging to the same semantic area, are highly related to each other. We also capture the structure topology and affinity among different regions of the input image. The specific approach is: 1) we divide the feature maps to different regions according to the coarse segmentation prediction, and then extract region-level features respectively. 2) we build a bi-directional graph, in which nodes represent regional features, and edges represent the affinity between two connected nodes. The bi-directional graph is used in region affinity reasoning. 3) The affinity-aware nodes are applied back to corresponding regions of the image, which help model the region dependencies and mitigate the unrealistic results. Additionally, a semantic region loss is proposed and employed to boost the pixel correlation and motivate the network to learn discriminative region-level and class-specific features. With the proposed approach, we achieve new state-of-the-art segmentation results on PASCAL-Context, ADE20K, and COCO-Stuff.

Figure 5: Qualitative segmentation examples.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561*, 2015.

[4] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. Pixelnet: Towards a general pixel-level architecture. *arXiv preprint arXiv:1609.06694*, 2016.

[5] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *CVPR*, 2016.

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018.

[7] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 2016.

[8] 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. *arXiv:1606.00915*, 2016.

[9] Wanli Chen, Xinge Zhu, Ruoqi Sun, Junjun He, Ruiyu Li, Xiaoyong Shen, and Bei Yu. Tensor low-rank reconstruction for semantic segmentation. *arXiv preprint arXiv:2008.00490*, 2020.

[10] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[11] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[12] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[13] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In *European Conference on Computer Vision*, pages 417–435. Springer, 2020.

[14] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6819–6829, 2019.

[15] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] 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 Conference on Computer Vision and Pattern Recognition*, 2019.

[17] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic segmentation with context encoding and multi-path decoding. *IEEE Transactions on Image Processing*, 29:3520–3533, 2020.

[18] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[19] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2), 2010.

[20] 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 Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[21] 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 international conference on computer vision*, pages 6748–6757, 2019.

[22] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *International Conference on Computer Vision*, pages 1–8. IEEE, 2009.

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

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

[25] 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*, 2016.

[26] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, pages 4003–4012, 2020.

[27] Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *ECCV*, 2020.

[28] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[29] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe L Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *IEEE International Conference on Computer Vision*, 2017.

[30] Jyh-Jing Hwang, Tsung-Wei Ke, Jianbo Shi, and Stella X Yu. Adversarial structure matching for structured prediction tasks. In *CVPR*, 2019.

[31] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, 2019.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[33] Md Amirul Islam, Shujon Naha, Mrigank Rochan, Neil Bruce, and Yang Wang. Label refinement network for coarse-to-fine semantic segmentation. *arXiv preprint arXiv:1703.00551*, 2017.

[34] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, 2018.

[35] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[36] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, 2011.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[38] M Pawan Kumar and Daphne Koller. Efficiently selecting regions for scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3217–3224. IEEE, 2010.

[39] Diane Larlus and Frédéric Jurie. Combining appearance models and markov random fields for category level object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[40] 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.

[41] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *European Conference on Computer Vision*, 2020.

[42] 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 International Conference on Computer Vision*, pages 9167–9176, 2019.

[43] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[45] Chang Liu, Henghui Ding, and Xudong Jiang. Towards enhancing fine-grained details for image matting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 385–393, 2021.

[46] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex C Kot. Feature boosting network for 3d pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):494–501, 2019.

[47] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems*, pages 1520–1530, 2017.

[48] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *IEEE International Conference on Computer Vision*, 2015.

[49] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[50] Jianhan Mei, Ziming Wu, Xiang Chen, Yu Qiao, Henghui Ding, and Xudong Jiang. Deepdeblur: text image recovery from blur to sharp. *Multimedia tools and applications*, 78(13):18869–18885, 2019.

[51] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[52] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision*, 2015.

[53] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 2015.

[56] Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE*

transactions on pattern analysis and machine intelligence, 2016.

[57] Bing Shuai, Henghui Ding, Ting Liu, Gang Wang, and Xudong Jiang. Toward achieving robust low-level and high-level scene parsing. *IEEE Transactions on Image Processing*, 28(3):1378–1390, 2018.

[58] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[59] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[60] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In *CVPR*, pages 3126–3135, 2019.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

[62] Suchen Wang, Yap-Peng Tan, Henghui Ding, Kim-Hui Yap, Junsong Yuan, and Ji-Yan Wu. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[63] Xiaohong Wang, Henghui Ding, and Xudong Jiang. Dermoscopic image segmentation through the enhanced high-level parsing and class weighted loss. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 245–249. IEEE, 2019.

[64] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[65] Xiaohong Wang, Xudong Jiang, Henghui Ding, and Jun Liu. Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation. *IEEE transactions on image processing*, 29:3039–3051, 2019.

[66] Xiaohong Wang, Xudong Jiang, Henghui Ding, Yuqian Zhao, and Jun Liu. Knowledge-aware deep framework for collaborative skin lesion segmentation and melanoma recognition. *Pattern Recognition*, page 108075, 2021.

[67] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv:1605.06885*, 2016.

[68] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018.

[69] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020.

[70] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1857–1866, 2018.

[71] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.

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

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

[74] 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 International Conference on Computer Vision*, pages 6798–6807, 2019.

[75] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018.

[76] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[77] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[78] 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.

[79] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–557, 2019.

[80] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2031–2039, 2017.

[81] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[82] 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 *European Conference on Computer Vision*, pages 267–283, 2018.

[83] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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