Attention-guided Chained Context Aggregation for Semantic Segmentation

Quan Tang Fagui Liu Jun Jiang Yu Zhang

South China University of Technology, Guangzhou, China

{csquantang, csjun.jiang, csemoszy}@mail.scut.edu.cn, fgliu@scut.edu.cn

Abstract

Recent breakthroughs in semantic segmentation methods based on Fully Convolutional Networks (FCNs) have aroused great research interest. One of the critical issues is how to aggregate multi-scale contextual information effectively to obtain reliable results. To address this problem, we propose a novel paradigm called the Chained Context Aggregation Module (CAM). CAM gains features of various spatial scales through chain-connected ladder-style information flows. The features are then guided by Flow Guidance Connections to interact and fuse in a two-stage process, which we refer to as pre-fusion and re-fusion. We further adopt attention models in CAM to productively recombine and select those fused features to refine performance. Based on these developments, we construct the Chained Context Aggregation Network (CANet), which employs a two-step decoder to recover precise spatial details of prediction maps. We conduct extensive experiments on three challenging datasets, including Pascal VOC 2012, CamVid and SUN-RGBD. Results evidence that our CANet achieves state-of-the-art performance. Codes will be available on the publication of this paper.

1. Introduction

Semantic segmentation is a vital task in computer vision, aiming to assign corresponding semantic labels to each pixel in images. It has fundamental applications in the fields of automatic driving [9, 44], medical image [45], augmented reality, *etc*. Fully Convolutional Networks (FCNs) [40], originating from deep convolutional neural networks (DCNNs) [32, 30, 21, 49, 52, 24] for image classification, achieve optimal performance in this task. They produce dense predictions by replacing fully connected layers with convolutions. FCNs [40] gain increasing receptive field and high-level contexts through cascaded convolutional and pooling layers. However, the continuous downsampling process causes the loss of spatial details, resulting in poor object delineation and small spurious regions. Figure 1 presents some examples. In summary, the paradox between semantics and spatial details is a significant predicament of DCNN approaches.

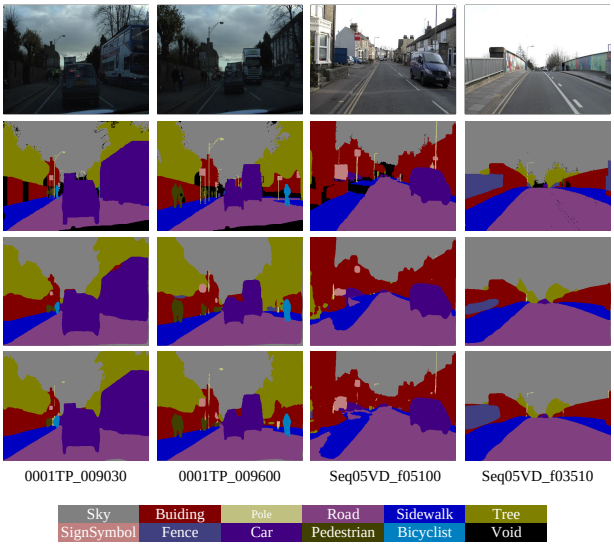


Figure 1. Some visualized predictions on CamVid test set. **First row**: input images. **Second row**: ground truth. **Third row**: predictions of FCN (baseline). **Fourth row**: predictions of CANet (ours). Both FCN and CANet are based on ResNet50. Because of the capability to capture multi-scale contextual information, CANet is able to ameliorate poor object delineation and small spurious regions. For example, in image "0001TP_009030" and "Seq05VD_f03510", FCN mistakes some pixels of cars for buildings and completely confuses fences with buildings, respectively. Meanwhile, taking advantage of the decoder, CANet obtains more sharper segmentation boundaries such as poles.

Combining dilated/atrous convolutions [4, 63] and context modules [68, 20, 66, 5, 6, 7, 65] becomes a popular alternative to reconcile the above contradiction. The dilated convolution can increase the receptive field while maintaining feature maps' resolutions without extra parameters. However, it suffers from the gridding dilemma [56], relinquishing part of the neighboring information that is also es-

sential for elaborate semantic segmentation on account of all pixels interacting with surrounding ones to make up objects and form local contexts. Context modules remedy this problem in some way. They are designed to join feature maps of various but larger receptive field to exploit both local and global contexts. Global cues help to understand the entire image scene and to some extent, reject the ambiguity caused by similar local objects, *e.g*. cars instead of ships are more likely to show up in a city scene.

Consequently, semantic segmentation needs to aggregate multi-scale features and balance local contexts and global cues [17], which is beneficial to segmenting objects of inconsistent spatial scales, like pedestrians and buildings in Figure 1. Most existing methods [68, 5, 6, 7, 20] adopt a parallel context module design that encodes contextual information through separate convolutional paths and fuse them in a specific stage at a time. Besides, stacked encoder-decoder structures [15, 41] are also employed to exploit contextual information, which can be considered as an in-series context module where the ladder features exclusively depend on the previous. HRNet [51] is a further developed network and provides multi-scale context aggregation at high resolution.

To enhance highly flexible aggregation of multi-scale contextual information, we propose a novel paradigm termed the Chained Context Aggregation Module (CAM), as illustrated with the red dashed box in Figure 2. CAM combines the advantages of in-series and parallel context modules. More specifically, CAM captures features of various scales at different levels and aggregates them in stages by the chain-connected ladder-style information flows. The Global Flow (GF) makes use of the shared features encoded by the backbone network to obtain global receptive field, which is advantageous to establish a perception of the entire image scene and reduce mislabeled pixels of similar objects. Differently, inputs of the Context Flow (CF) consists of two parts: the shared features and the output features of upper flow. Stacked convolutional and pooling layers with various strides are adopted within CFs to exploit local contexts of individual scales. We name the process *pre-fusion*. Different information flows obtain feature maps of different spatial scales and they are then aggregated by Residual Connections, which we name as *re-fusion*. We further apply attention models to perform recombination and selection on the multi-scale features for refining the segmentation results. After the two-stage context aggregation, the final fused feature map encodes rich contextual information that is deciding for accurate segmentation. Besides, we adopt a simple yet capable two-step decoder to recover more spatial details during upsampling, as shown by the blue dashed box in Figure 2. It enhances the semantics of low-level features without spatial loss by the adapted Global Convolutional Network (AGCN) and makes them harmoniously fuse with high-level contexts. Finally, we construct the Chained Context Aggregation Network (CANet) for semantic image segmentation and conduct extensive experiments on three challenging datasets whose results demonstrate the validity.

We conclude the critical contributions as follows:

- We propose the Chained Context Aggregation Module (CAM) to enable sufficiently flexible and powerful aggregation of multi-scale contexts in a two-stage process, which remarkably improves the segmentation performance.
- We further utilize attention models in CAM to fuse those multi-scale features efficiently and refine the results.
- A simple decoder is manipulated to restore fine spatial details during upsampling.
- We construct a generalized framework termed the Chained Context Aggregation Network (CANet) and achieve impressive results on the benchmarks of Pascal VOC 2012, CamVid and SUN-RGBD.

2. Related Work

With increasing applications of deep learning methods to semantic segmentation in recent years, the task has made breakthroughs on benchmarks. We briefly review related research works in this section.

Encoder-decoder. Encoder-decoder structures have been successfully applied in semantic segmentation. Typically, they contain an encoder that downsamples the input image to obtain high-level semantics and a decoder that gradually restores the resolution to classify every pixel. Both SegNet [2] and UNet [45] utilize a symmetric decoder to obtain fine-recovered predictions. GCN [43] and RefineNet [35] are further developments with carefully designed decoders. SDN [15] exploits contextual information by stacking multiple encoder-decoders. Deconvolution [42] and DUpsample [53] employ upsampling strategies different from bilinear interpolation as decoders to achieve better results.

Spatial information. FCN [40] based methods obtain high-level semantics through downsampling operations at the expense of spatial details. Yu and Koltun [63] develop dilated convolutions to reduce spatial loss during encoding and achieve excellent results. Following this idea, PSPNet [68], Deeplab v3+ [7], APCNet [20], DANet [13] and CFNet [66] all apply dilated convolutions in the backbone network to maintain the resolution of feature maps. FRRN [44] and HRNet [51] employ residual streams to serve the purpose. Another idea [43] is to utilize large-kernel convolutions to increase the receptive field quickly.

Multi-scale context. At the top of the backbone network, PSPNet [68] and Deeplab [5, 6, 7] employ parallel information flows to perceive multi-scale features by different pooling strides and dilated rates respectively. APCNet [20], DANet [13] and CFNet [66] take advantage of attention models to obtain various context information. Based on the U-shape structure, RefineNet [35], LFNet [62] and GCN [43] achieve a productive fusion of hierarchical features. DFANet [33] implements an in-depth aggregation of hierarchical features by cascading multiple encoders. Besides, Recurrent Neural Networks (RNNs) are also developed to capture long-range dependencies [47, 12].

Attention mechanism. The core idea of attention mechanism is to assign distinctive attention weights to different parts of the input, just like people focusing on attractive parts of the input features. Like the non-local block [57] introducing self-attention [54] into computer vision, various types of attention mechanisms [26, 1, 23, 34, 59, 16] play an increasingly important role in this field. Methods with attention models for semantic segmentation [20, 65, 13, 12, 64, 25] also further improve the performance.

3. Method

We propose the Chained Context Aggregation Network (CANet) to enable flexible capturing and aggregation of multi-scale contextual information and to explore its improvement for semantic segmentation. We elaborate on this network in this section.

3.1. Network Overview

Our Chained Context Aggregation Network (CANet) adopts an asymmetric encoder-decoder architecture. The backbone network employs dilated ResNet [21, 5, 68] and the output feature maps $\{\boldsymbol{C}_i\}_{i=1,2,3,4}$ of each residual module have strides $\{4, 8, 8, 8\}$ respectively, compared with the input image. Figure 2 depicts the overview of CANet. In particular, following He *et al*. [22], we replace the original 7×7 convolution with three stacked 3×3 convolutional layers. Here supplanting the ResNet with other networks is straightforward. At the top of the backbone lies the carefully designed Chained Context Aggregation Module (CAM), expecting to exploit multi-scale contextual features effectively and aggregate them flexibly to improve segmentation performance for objects of various spatial scales. Finally, a two-step decoder is adopted to upsample the prediction maps to present the classification confidence for each class at every pixel. CANet utilizes bilinear interpolation as a naive upsampling strategy.

3.2. Chained Context Aggregation Module

The Chained Context Aggregation Module (CAM), as illustrated by the red dashed box in Figure 2, is the critical part of CANet to aggregate multi-scale contextual information. Based on the shared features encoded by the backbone network, CAM further exploits semantic relations of different spatial scales through Global Flow (GF) and Context Flows (CFs). They are joined through Flow Guidance Connections to interact and fuse. CF firstly *pre-fuse* two different-scale features coming from the backbone and upper information flow. CAM then *re-fuse* features of various scales obtained by GF and CFs under the guidance of Residual Connections to enrich contextual information and serve for precise predictions of each pixel's label. We can see that during the above process, the context obtained by a lower information flow does not entirely depend on that of the upper. It is the two-stage feature exploitation that enables flexible multi-scale context aggregation. Besides, we employ attention models to re-weigh the features at spatial and channel domains sequentially to promote the future aggregation and refine the final segmentation results. The main components of CAM are described below.

Global Flow. Practices [68, 6, 66, 38] have witnessed that global pooling feature can provide global receptive field as a reliable cue to distinguish confusing objects. In CAM, it is obtained by applying global average pooling on the shared feature map of the backbone network, which we refer to as the Global Flow (GF). Figure 3(a) depicts its details.

Context Flow. We propose Context Flows (CFs) to exploit local contexts of different receptive field as shown in Figure 3(c). CF is the spotlight where *pre-fusion* occurs. Inputs of a CF are composed of the shared features obtained by the backbone and the features of the upper information flow. We devise a Context Fusion Module (CFM) to fuse them effectively. More specifically, given the two inputs of different spatial scales, CFM firstly concats them in the channel dimension and downsampling N times through an average pooling layer that also lessens computation cost. The following two consecutive convolutional layers can eliminate the aliasing effect and further enlarge the receptive field, where group convolutions and channel shuffle [67] are adopted to diminish model parameters. Then the spatial-attention based Context Refinement Module (CRM) refines the fused features, which is specified in the later subsection. Finally, CF upsamples the output to the same size of the input. Note that we set the number of output feature maps of the two group convolutions to 768.

Flow Guidance Connections. GF and CFs with different downsampling rates are proposed to obtain global and local contexts of various spatial scales respectively. We propose

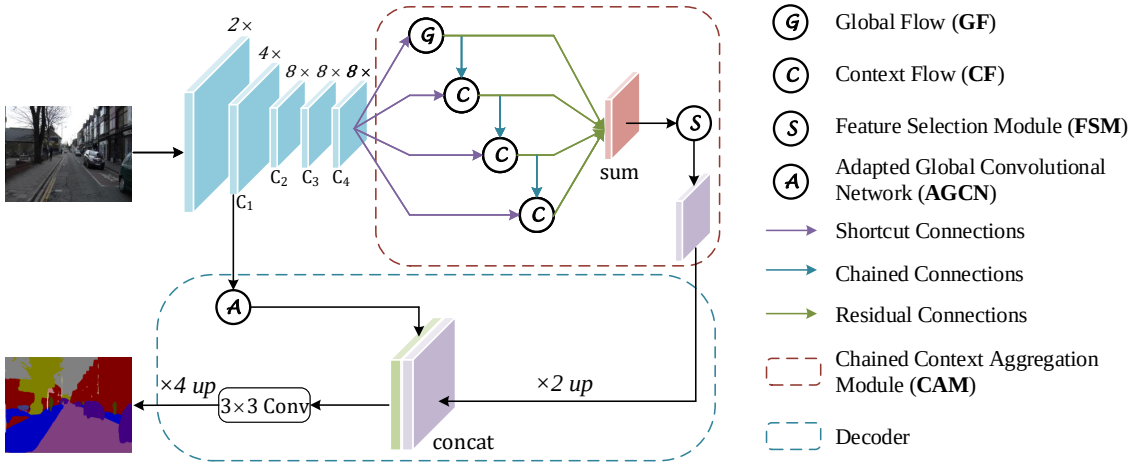


Figure 2. Overview of our proposed CANet. Given an input image, we first adopt a DCNN to encode a shared feature map, then the carefully designed Chained Context Aggregation Module (CAM) is applied to enrich multi-scale contexts, followed by a two-step decoder to get the final per-pixel prediction. "$N\times$" indicates the output stride of the feature map and "$\times N$ up" means N-time upsampling operation. Shortcut connections, Chained Connections and Residual Connections are collectively called Flow Guidance Connections.

Flow Guidance Connections to unite GF and CFs to enhance feature delivering and interaction, and enrich multi-scale contextual information. Specifically, Flow Guidance Connections include Shortcut Connections, Chained Connections and Residual Connections, as separately depicted by the purple, blue and green arrows in Figure 2. Shortcut Connections let CAM reuse the output features of the backbone network. They not only efficiently promote the acquisition of multi-scale contexts but also reduce model parameters. Since GF and CFs have distinct receptive field, Chained Connections are designed to guide the *pre-fusion* and magnify the flexible feature fusion between them. Finally, Residual Connections serve as ushers of the *re-fusion* process. It also alleviates gradient vanishing caused by the CAM increasing the network's depth, which is contributory to model convergence. The *re-fused* feature map is then fed into Feature Selection Module (FSM) to perform recombination and selection of advantageous features.

Different combinations of different quantity and downsampling rates of CFs make up diverse CAMs to exploite multi-scale contexts. It is another embodiment of the flexibility of CAM besides the two-stage aggregation mechanism. Figure 4(a) provides a possible combination where the information flow symbolized by the black dotted arrow makes up a stack of multiple shallow encoder-decoders that can be regarded as a particular case of SDN [15].

3.3. Attention-guided CAM

From the above elaboration, we can grasp that GF and CFs are capable of encoding a variety of contexts. Although diverse contextual information is beneficial for delicate semantic segmentation, it also poses challenges for feature aggregation. Therefore, we adopt a spatial-attention based Context Refinement Module (CRM) and channel-attention [23, 34] based Feature Selection Module (FSM) to facilitate the *pre-fusion* and *re-fusion* process in CAM, respectively. Within a CF, CRM re-weighs the multi-scale feature of every pixel to improve the representation of con-

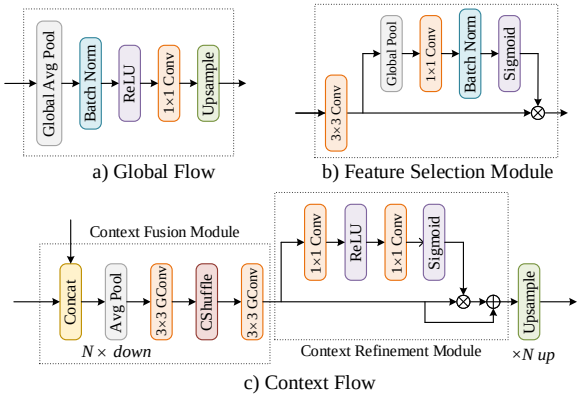


Figure 3. Detailed components of CAM. "$N\times$ down" means N-time downsampling of the inputs by average pooling layer, while "$\times N$ up" means N-time upsampling operation. GConv and CShuffle are shorthand for Group Convolutions and Channel Shuffle respectively. $\oplus$ and $\otimes$ represents element-wise summation and element-wise product, respectively.

textual information at the spatial level, promoting the feature *pre-fusion*. Meanwhile, features for *re-fusion* contain a variety of contextual scales, so we propose FSM to magnify advantageous features and suppress the useless or harmful at the channel level. Both attention models not only promote the fusion of features of different scales but also help to balance their effect on the segmentation results. Experiments in Section 4 demonstrate their refinement on final predictions.

Context Refinement Module (CRM). Assume the output feature map of CFM in the ith CF is $\boldsymbol{X}_i$, and $\{x_{i,j}^q\}_{q=1,2...,Q}$ represents the pixel feature of the qth feature map at position j. Note that Q is a constant for all $\boldsymbol{X}_i$, which is empirically set to 768 in this paper. We re-weigh the original pixel feature as

$$z_{i,j}^q = x_{i,j}^q + r_{i,j}^q \cdot x_{i,j}^q \tag{1}$$

where $z_{i,j}^q$ is the re-weighted feature of the qth feature map at position j. All $z_{i,j}^q$ make up the re-weighted feature map $\boldsymbol{Z}_i$. $r_{i,j}^q \in (0,1)$ is the corresponding attention weight for each pixel. For different images, $r_{i,j}^q$ should have different values that characterize the importance of the pixel feature at the corresponding position. In other words, it should be calculated by learnable parameters. Let $\boldsymbol{R}_i$ represent all attention weights, and we calculate it by

$$\boldsymbol{R}_i = F_{spatial}(\boldsymbol{X}_i, \boldsymbol{W}) = \sigma(f(\delta(f(\boldsymbol{X}_i, \boldsymbol{W}_1)), \boldsymbol{W}_2)) \tag{2}$$

where $F_{spatial}(\cdot,\cdot)$ is the spatial-attention function and $f(\cdot,\cdot)$ denotes convolution. σ and δ refer to the Sigmoid and ReLU function respectively. $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{Q}{S} \times Q}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{Q \times \frac{Q}{S}}$ are both learnable weights, where s is set to 16 for reducing computation cost and model parameters. We implement this formulated spatial-attention model by the structure shown in Figure 3(c). The bottleneck composed of two 1×1 convolutional layers computes the attention weights that re-weigh features of each pixel. The cooperation between CRM and CFM promotes the *pre-fusion* of two different-scale features that are concurrently fed into a CF. The re-weighted features are also beneficial to the *re-fusion* of features encoded by different information flows.

Feature Selection Module (FSM). Let $\boldsymbol{G}$ be the global pooling feature obtained by GF, and $\boldsymbol{U}$ the aggregation feature map of GF and CFs, then

$$\boldsymbol{U} = F_{bilinear}(\boldsymbol{G}) + \sum_i F_{bilinear}(\boldsymbol{Z}_i) \tag{3}$$

where $F_{bilinear}$ represents the bilinear interpolation function. The resulting $\boldsymbol{U}$ aggregates rich semantic information of multiple scales. They are then effectively recombined and selected by FSM, whose details are illustrated in Figure 3(b). During forward propagation, we first project $\boldsymbol{U}$ by a 3×3 convolution to eliminate the aliasing effect and further expand the receptive field. Then a global average pooling layer and a 1×1 convolution are applied to compute the attention weights on channel dimension, which contributes to amplifying the advantageous features of $\boldsymbol{U}$. The 3×3 convolution contains an activation layer and a normalization layer unifying feature scales and helping to obtain attention scores of the same scale. We formulate the above process as follows:

$$\boldsymbol{U}' = f(\delta(F_{BN}(\boldsymbol{U})), \boldsymbol{W}_3) \tag{4}$$

$$\widetilde{\boldsymbol{U}} = \boldsymbol{U}' \otimes \sigma(F_{BN}(f(F_{GAP}(\boldsymbol{U}'), \boldsymbol{W}_4))) \tag{5}$$

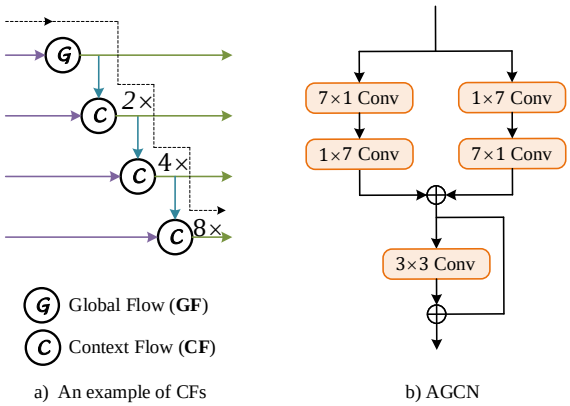
where F_{BN} and F_{GAP} represent the normalization layer and global average pooling layer, respectively. $\boldsymbol{W}_3$ and $\boldsymbol{W}_4$ are both learnable parameters. $\otimes$ represents element-wise product. $\widetilde{\boldsymbol{U}}$ is then upsampled by the decoder to obtain predictions.

3.4. Decoder

In convolutional networks, each layer handles diverse information. Low-level layers usually have more positional information and high-level ones hold more semantics. Both positions and semantics play a pivotal role in semantic segmentation. Therefore, this paper employs a simple two-step decoder to explore the semantic relations between high- and low-level features, and further fuse them to provide abundant semantic and positional clues for semantic segmentation. During decoding, the output feature map of CAM is first upsampled two times to blend with low-level features of the same resolution and then upsampled four times to the resolution of the input image using bilinear interpolation. Deeplab v3+ [7] adopts just a 1×1 convolution to project low-level features. However, inspired by the idea of large-kernel, we introduce an adapted Global Convolutional Network, which we refer to as AGCN, to polish semantic expressiveness without loss of resolution. It enables low-level positions harmoniously fuse with high-level semantics. Figure 4(b) illustrates the AGCN design. Refer to [43] for more details.

4. Experimental Results

We conduct extensive experiments on Pascal VOC 2012 [11], CamVid [3], and SUN-RGBD [50] to evaluate the performance of our proposed CANet. All results are obtained with multi-scale and flipping skills if not specified. We adopt the standard benchmarks, pixel accuracy (PA) and mean intersection over union (mIoU), as the evaluation metrics. We use only mIoU on Pascal VOC 2012 for the common convention.



a) An example of CFs b) AGCN

Figure 4. a) One possible combination of GF and CFs. "$N\times$" indicates the down-sampling rate of the pooling layer in CFs. b) Skip connection in decoder that employs large-kernel convolutions adapted from Global Convolutional Network, which we referred to as AGCN.$\oplus$ is element-wise summation.

Backbone	CFs	mIoU(%)
ResNet50	None(baseline)	70.48
ResNet50	{2}	78.47
ResNet50	{2, 2}	78.09
ResNet50	{2, 4}	78.06
ResNet50	{2, 2, 2}	78.26
ResNet50	{2, 4, 4}	78.36
ResNet50	{2, 4, 8}	78.03
ResNet50	{2, 2, 2, 2}	**79.17**
ResNet50	{2, 4, 8, 12}	78.01
ResNet50	{2, 2, 2, 2, 2}	77.94
ResNet101	{2, 2, 2, 2}	**81.57**

Table 1. Investigation of the quantity and down-sample rates of CFs. $\{d_1, d_2, , d_n\}$ means there are n CFs chain-connected from top down, and the downsampling rates are $d_1, d_2, ..., d_n$ respectively. For fair comparison, baseline here adopts auxiliary loss using the $\boldsymbol{C}_3$ feature map whose balance coefficient is set to 0.5.

4.1. Implementation Details

Our experiments are based on the deep learning framework Apache MXNet [8] and we borrow the structure and ImageNet [46] pre-trained model parameters of the backbone ResNet from the open-source toolkit GluonCV [22, 18]. As mentioned above, we set the dilate rates of backbone's last two residual modules to 2 and 4 respectively. Thus the resolution of backbone's final output feature map is 1/8 of the input image. A two-step bilinear interpolation decoder recovers the resolution for predicting semantic labels of each pixel. Following prior works [20, 66], we use the poly learning rate policy $lr = baselr \times (1 - \frac{iter}{total\_iter})^{power}$ and set the power to 0.9. The initial learning rate is 0.001 for all three datasets. We use the standard mini-batch stochastic gradient descent (SGD) as the optimizer and set momentum to 0.9. To prevent over-fitting, we set the weight decay to 1e-4 for Pascal VOC 2012 and SUN-RGBD, and 1e-3 for CamVid. For data augmentation, we first flip input images with a probability of 0.5 and randomly scale them from 0.5 to 2.0 times. Then we crop the images with padding if needed. Finally, a random Gaussian blur is added. Since appropriate crop size influences the model performance, we empirically crop images to 512×512 for Pascal VOC 2012 and SUN-RGBD, and 360×360 for CamVid.

Since CAM and the decoder further deepen CANet, here we employ two additional auxiliary loss functions to better supervise the network training, namely deep supervision [55]. The joint loss function is defined as follows:

$$L = L_p + \lambda_1 L_u + \lambda_2 L_c \tag{6}$$

where L_p represents the principal loss to supervised the entire network. L_u and L_c are two auxiliary losses, which are calculated by the aggregation feature map $\boldsymbol{U}$ described in Section 3.3 and $\boldsymbol{C}_3$ in Section 3.1, respectively. They mainly supervise the training of CAM and the backbone network. All loss functions are pixel-wise softmax cross-entropy loss. λ_1 and λ_2 are adopted to balance the training process, and we set both to 0.5. We do not use the auxiliary outputs when inference.

4.2. Results on Pascal VOC 2012

Pascal VOC 2012 [11] contains 1464, 1449 and 1456 images for training, validation and testing, respectively. All images are pixel-wise labeled with 21 semantic classes, one of which is background. Following prior works, we augment the training set with SBD dataset [19] for experiments, resulting in 10582 images for training.

We first perform sound ablation studies on Pascal VOC 2012 validation set to evaluate the benefits of key components in CANet as well as to explore the improvement of different combinations of GF and CFs on the segmentation results. Our baseline is dilated ResNet based FCN [40, 21, 5]. All ablation results are based on single scale inputs and without MS-COCO [37] pre-training and fine-tuning on the original dataset.

Ablation for the quantity and down-sampling rates of CFs. We believe that different quantities and down-sampling rates of CFs contribute to capture different contextual information of objects with various scales. Figure 4(a) gives a possible example of CFs. We conduct some exploratory experiments on this and Table 1 reports the results. It can be seen that: 1) Compared to the baseline FCN,

Ablation Components	mIoU(%)
None(baseline)	69.97
AL(baseline)	70.48
GF&CFs	78.58
GF&CFs+CRM	78.69
GF&CFs+FSM	78.72
GF&CFs+CRM+FSM	78.80
GF&CFs+CRM+FSM+Decoder(1x1 Conv)	78.94
GF&CFs+CRM+FSM+Decoder(AGCN)	79.03
GF&CFs+CRM+FSM+Decoder(AGCN)+AL	**79.17**

Table 2. Ablation results of key components and auxiliary loss (AL) stated in Sec. 4.1. The backbone network is ResNet50.

the segmentation performance is greatly improved no matter what combination of CFs is. And the $\{2,2,2,2\}$ achieves the best with an 8.69% mIoU improvement. 2) The increase of CF's quantity within the range of $1 \sim 4$ brings more multi-scale features hence improves the performance. However, the performance degrades noticeably when the number goes beyond 4. 3) Different combinations of CFs introduce performance perturbation. 4) A deeper backbone network further enhances the performance. We employ the $\{2,2,2,2\}$ in the following experiments.

Ablation for key components and auxiliary loss. We show the effectiveness of key components in CANet by adding them on baseline one by one. The experimental results are listed in Table 2. We can see that: 1) The chained-connected ladder-style information flows, *i.e*. GF&CFs, can significantly improve the semantic segmentation performance, from mIoU 69.97% to 78.58%. 2) Attention models CRM and FSM proffer minor enhancement. When they work together, the enhancement enlarges. 3) The decoder brings more spatial details, thus refine the results. 4) For both baseline and CANet, the use of auxiliary loss is beneficial to the model convergence.

Ablation for Global Flow. We validate the benefits of GF in obtaining global receptive field and rejecting local ambiguities based on different backbone networks. The uppermost CF takes only the shared features as input when there is no GF. Table 3 shows the results and Figure 5 visualizes some comparisons. Results indicate that GF has a noticeable improvement in semantic segmentation performance, which is beneficial to subduing misclassification of similar local pixels.

Comparisons with State-of-the-Arts. We conduct experiments on the testing set to compare with other SOTA methods. We adopt the ImageNet [30] pre-trained ResNet101 as the backbone. CANet is first pre-trained on the augmented dataset, then fine-tuned with the original *trainval* images. Results based on multi-scale and flipping testing skills are

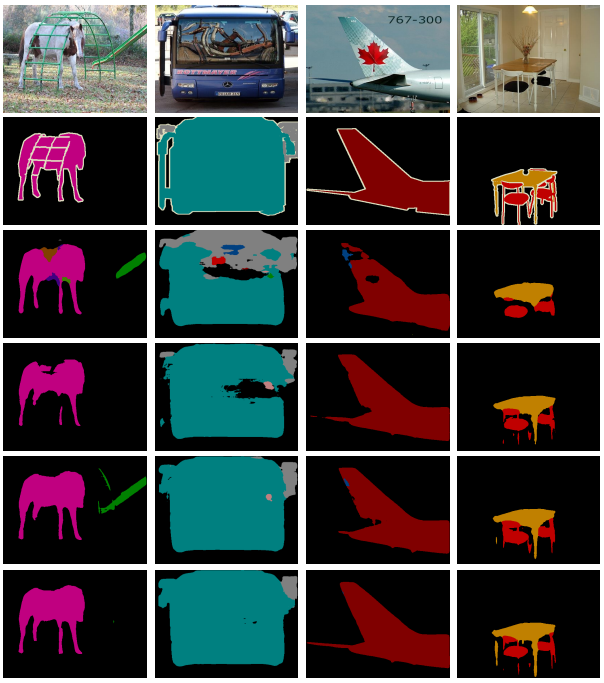


Figure 5. Some visualized comparisons with baseline FCN on Pascal VOC 2012 validation set. **First row**: input images. **Second row**: ground truth. **Third row**: predictions of FCN (baseline). **Fourth row**: predictions of CANet without GF. **Fifth row**: predictions of CANet without attention models, *i.e*. CRM and FSM. **Sixth row**: predictions of CANet. Both FCN and CANet are based on ResNet50. It can be noticed that CANet introduced much less mislabeled pixels and get more sharper segmentation boundaries. Due to the ability of full scene understanding and selection of beneficial features respectively, GF and attention models help to eliminate local ambiguity.

Backbone	GF	mIoU(%)
ResNet50	×	77.90
ResNet50	✓	**79.17**
ResNet101	×	80.32
ResNet101	✓	**81.57**

Table 3. Investigation with different backbones on the importance of Global Flow which captures global pooling features.

reported in Table 4. Our CANet achieves mIoU 84.7% and outperforms all existing approaches. Additionally, when using MS-COCO [37] pre-training, we obtain mIoU 87.2%.

4.3. Results on CamVid

CamVid [3] is a street scene dataset that contains both light and dark conditions. The distinction between CamVid and Pascal VOC 2012 [11] lies in the former's more complicated scenes and more substantial inconsistency in spa-

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU(%)
FCN [40]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLabv2 [5]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
DeconvNet [42]	89.9	39.9	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
DPN [39]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [36]	90.6	37.6	80.0	67.8	74.4	92.0	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
ResNet38 [60]	94.4	72.9	94.9	68.8	78.4	90.6	90.0	92.1	40.1	90.4	71.7	89.9	93.7	91.0	89.1	71.3	90.7	61.3	87.7	78.1	82.5
PSPNet [68]	91.8	71.9	94.7	71.2	75.8	95.2	89.9	95.9	39.3	90.7	71.7	90.5	94.5	88.8	89.6	72.8	89.6	**64.0**	85.1	76.3	82.6
EncNet [65]	94.1	69.2	**96.3**	**76.7**	**86.2**	96.3	90.7	94.2	38.8	90.7	73.3	90.0	92.5	88.8	87.9	68.7	92.6	59.0	86.4	73.4	82.9
SDN [15]	**96.2**	73.9	94.0	74.1	76.1	96.7	89.9	**96.2**	44.1	92.6	72.3	91.2	94.1	89.2	89.7	71.2	93.0	59.0	88.4	76.5	83.5
CFNet [66]	95.7	71.9	95.0	76.3	82.8	94.8	90.0	95.9	37.1	92.6	73.0	**93.4**	94.6	89.6	88.4	74.9	**95.2**	63.2	**89.7**	78.2	84.2
APCNet [20]	95.8	**75.8**	84.5	76.0	80.6	**96.9**	90.0	96.0	42.0	93.7	**75.4**	91.6	95.0	90.5	89.3	**75.8**	92.8	61.9	88.9	79.6	84.2
CANet(ours)	95.0	73.0	95.5	76.5	79.8	94.0	**93.8**	95.7	**46.9**	**94.9**	69.5	92.9	**95.8**	**92.6**	**91.5**	73.1	92.8	61.5	87.2	**81.5**	**84.7**
CANet(ours)†	96.3	69.3	96.7	80.7	84.0	97.7	94.0	97.0	49.3	95.5	80.4	95.3	96.3	92.8	91.2	78.8	94.9	73.5	89.9	80.1	**87.2**

Table 4. Per-class scores on Pascal VOC 2012 testing set. '†' indicates MS-COCO pre-training.

Method	PA(%)	mIoU(%)
SegNet [2]	62.5	46.4
DeconvNet [42]	85.6	48.9
Bayesian SegNet [29]	86.9	63.1
Dilation8 [63]	79.0	65.3
HDCNN-448+TL [58]	90.9	65.6
Dilation8+FSO [31]	88.3	66.1
FC-DenseNet103 [28]	91.5	66.9
DCDN [14]	91.4	68.4
SDN [15]	91.7	69.6
SDN+ [15]	92.7	71.8
CANet-ResNet50(ours)	93.1	73.3
CANet-ResNet101(ours)	**93.2**	**74.1**

Table 5. Experimental results on CamVid dataset (11 classes). Our CANet surpasses existing methods with a large margin.

Method	PA(%)	mIoU(%)
FCN [40]	68.2	27.4
DeconvNet [42]	66.1	22.6
Bayesian SegNet [29]	71.2	30.7
SegNet [2]	72.6	31.8
DeepLabv2 [5]	71.9	32.1
Piecewise [36]	78.4	42.3
RefineNet-ResNet101 [35]	80.4	45.7
Ding *et al*. [10]	81.4	47.1
CANet-ResNet101(ours)	**81.9**	**47.7**

Table 6. Quantitative results on SUN-RGBD dataset (37 classes) which only use RGB modality for evaluation.

tial scale of objects. Accordingly, CamVid needs a further robust model that captures rich multi-scale contextual information. We use the dataset described by SegNet [2] that contains 367 images for training, 100 images for validation, and 233 images for testing, all labeled with 11 semantic categories. We train CANet with training and validation images and report the performance on the testing images. Table 5 reports the results. Our CANet greatly surpasses the existing optimal SDN+ [15] by a 2.3% mIoU improvement. Note that unlike SDN+, we do not pre-train CANet on Pascal VOC 2012. Figure 1 gives some visualizations.

4.4. Results on SUN-RGBD

SUN-RGBD dataset [50] has a total of 10335 indoor images collected from NYU depth v2 [48], Berkeley B3DO [27] and SUN3D [61], of which 5280 images are for training and 5050 images for testing. It provides pixel-wise labeling for 37 semantic labels. There are various objects in one image scene and they differ in shapes, sizes and even spatial poses, which makes SUN-RGBD one of the most challenging datasets. In this paper, we only utilize RGB modality for experiments. Quantitative results reported in Table 6 demonstrate that our CANet achieves SOTA performance.

5. Conclusion

In this paper, we propose a new paradigm termed the Chained Context Aggregation Module (CAM) to sufficiently exploit multi-scale contexts for accurate semantic segmentation. It can effectively capture semantics of various spatial scales through *pre-fusion* and *re-fusion* of multiple information flows and give a remarkable improvement on the model performance. We further employ two attention models to promote the feature fusion, recombination and selection. A two-step decoder is employed to recover fine spatial details. Extensive experiments on three challenging datasets indicate the effectiveness and advancement of our CANet, which obtains mIoU 84.7% on Pascal VOC 2012 test set without MS-COCO pre-training and any post-processing. We hope this flexible feature aggregation paradigm can bring new vitality to the semantic segmentation community.

References

[1] Amjad Almahairi, Nicolas Ballas, Tim Cooijmans, Yin Zheng, Hugo Larochelle, and Aaron Courville. Dynamic capacity networks. In *International Conference on Machine Learning*, pages 2549–2558, 2016. 3

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2, 7, 8

[3] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 5, 7

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1

[5] 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. 1, 2, 3, 6, 8

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2, 3

[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. 1, 2, 3, 5

[8] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015. 5

[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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[10] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2393–2402, 2018. 8

[11] 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):303–338, 2010. 5, 6, 7

[12] Heng Fan, Xue Mei, Danil Prokhorov, and Haibin Ling. Multi-level contextual rnns with attention model for scene labeling. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3475–3485, 2018. 3

[13] 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. 2, 3

[14] Jun Fu, Jing Liu, Yuhang Wang, and Hanqing Lu. Densely connected deconvolutional network for semantic segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3085–3089. IEEE, 2017. 8

[15] Jun Fu, Jing Liu, Yuhang Wang, Jin Zhou, Changyong Wang, and Hanqing Lu. Stacked deconvolutional network for semantic segmentation. *IEEE Transactions on Image Processing*, 2019. 2, 4, 8

[16] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017. 3

[17] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 2

[18] Jian Guo, He He, Tong He, Leonard Lausen, Mu Li, Haibin Lin, Xingjian Shi, Chenguang Wang, Junyuan Xie, Sheng Zha, et al. Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing. *arXiv preprint arXiv:1907.04433*, 2019. 6

[19] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 6

[20] 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. 1, 2, 3, 6, 8

[21] 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. 1, 3, 6

[22] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019. 3, 6

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3, 4

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[25] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*, 2018. 3

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3

[27] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer depth cameras for computer vision*, pages 141–165. Springer, 2013. 8

[28] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017. 8

[29] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 8

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 7

[31] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3168–3175, 2016. 8

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[33] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019. 3

[34] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 4

[35] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 2, 3, 8

[36] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016. 8

[37] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 7

[38] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 3

[39] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE international conference on computer vision*, pages 1377–1385, 2015. 8

[40] 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. 1, 2, 6, 8

[41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[42] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2, 8

[43] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017. 2, 3, 5

[44] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4151–4160, 2017. 1, 2

[45] 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*, pages 234–241. Springer, 2015. 1, 2

[46] 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):211–252, 2015. 6

[47] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Dag-recurrent neural networks for scene labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3620–3629, 2016. 3

[48] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 8

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[50] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5, 8

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 2

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with

convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[53] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3126–3135, 2019. 2

[54] 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*, pages 5998–6008, 2017. 3

[55] Liwei Wang, Chen-Yu Lee, Zhuowen Tu, and Svetlana Lazebnik. Training deeper convolutional networks with deep supervision. *arXiv preprint arXiv:1505.02496*, 2015. 6

[56] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018. 1

[57] 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. 3

[58] Yuhang Wang, Jing Liu, Yong Li, Jun Fu, Min Xu, and Hanqing Lu. Hierarchically supervised deconvolutional network for semantic video segmentation. *Pattern Recognition*, 64:437–445, 2017. 8

[59] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 3

[60] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 8

[61] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1625–1632, 2013. 8

[62] 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. 3

[63] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1, 2, 8

[64] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 3

[65] 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. 1, 3, 8

[66] 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. 1, 2, 3, 6, 8

[67] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018. 3

[68] 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. 1, 2, 3, 8