Boundary-Aware Feature Propagation for Scene Segmentation

Henghui Ding[1]   Xudong Jiang[1]   Ai Qun Liu[1]   Nadia Magnenat Thalmann[1,2]   Gang Wang[3]

[1]Nanyang Technological University, Singapore

[2]University of Geneva, Switzerland   [3]Alibaba Group, China

{ding0093, exdjiang, eaqliu, nadiathalmann}@ntu.edu.sg, gangwang6@gmail.com

Abstract

In this work, we address the challenging issue of scene segmentation. To increase the feature similarity of the same object while keeping the feature discrimination of different objects, we explore to propagate information throughout the image under the control of objects' boundaries. To this end, we first propose to learn the boundary as an additional semantic class to enable the network to be aware of the boundary layout. Then, we propose unidirectional acyclic graphs (UAGs) to model the function of undirected cyclic graphs (UCGs), which structurize the image via building graphic pixel-by-pixel connections, in an efficient and effective way. Furthermore, we propose a boundary-aware feature propagation (BFP) module to harvest and propagate the local features within their regions isolated by the learned boundaries in the UAG-structured image. The proposed BFP is capable of splitting the feature propagation into a set of semantic groups via building strong connections among the same segment region but weak connections between different segment regions. Without bells and whistles, our approach achieves new state-of-the-art segmentation performance on three challenging semantic segmentation datasets, i.e., PASCAL-Context, CamVid, and Cityscapes.

1. Introduction

Scene segmentation is a challenging and fundamental task that aims to assign semantic categories to every pixels of scene images. The key of scene segmentation refers to parsing and segmenting a scene image into a range of semantic coherent regions. Therefore, it is critical to improve the feature similarity of the same object while keeping the feature discrimination of different objects. To this end, on the one hand, we explore to propagate features throughout the images to share features and harvest context information, which is beneficial for improving the feature similarity. One the other hand, in order to keep the discrim-
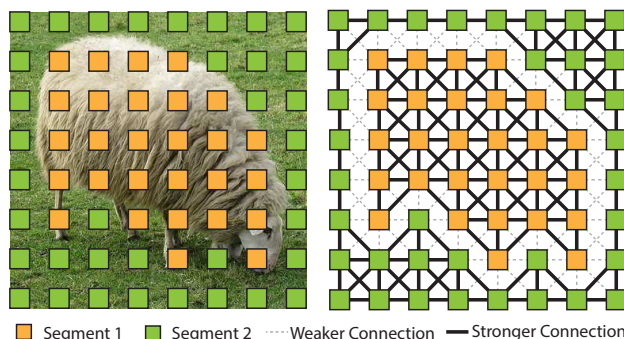

Figure 1. (Best viewed in color) The boundary-aware feature propagation module builds stronger connections within the same segment and weaker connections between different segments, which helps to enhance the similarity of features belonging to the same segment while keeping discrimination of features belonging to different segments.

inative power of features belonging to different objects, we propose to make use of boundary information to control the information flow during propagation progress. In a word, we propose a boundary-aware feature propagation module to build strong connections within the same segment and weak connections between different segments, as shown in Figure 1. This module requires two components: boundary detection and graph construction.

First, boundary detection, which is an implicit task in scene segmentation, is important for meticulous dense prediction. However, in existing segmentation methods, boundary detection did not attract due attention since boundary pixels only account for a small portion of the entire image and it has little contribution to the performance improvement. In this work, we try to find a way to simultaneously achieve segmentation and boundary detection, and further make use of the learned boundary information to enhance the segmentation performance. With regards to this, we propose to generate the boundary label of semantic objects from the existing object class labels given in the segmentation datasets and define it as an additional class for learning and classification. By doing so, concise boundaries

are well learnt and inferred as one additional class because the characteristics of pixels on boundary are different from those of most pixels off boundary. And the parsing of pixels among disputed area (i.e., near the boundary) is enhanced. Moreover, taking boundary as an additional class requires little change on network but makes the network be aware of the boundary layout that could be further used for segmentation improvement.

Second, graphic model is needed to create the order rules for feature propagation. Convolutional methods [13, 79] are popular in scene segmentation, but they usually consume large computation resources when aggregating features from grand range of receptive fields. Moreover, the convolution kernels could not vary with input resolutions, thus cannot ensure a holistic view of the overall image. Recently, DAG-RNN [66] proposes to use four directed acyclic graphs (DAGs) with different directions to model the function of undirected cyclic graphs (UCGs), which structurize the image by building pixel-by-pixel connections throughout whole image. However, DAGs require lots of loops to scan the images pixel by pixel. Thus it is very slowly even on low-resolution feature maps, which limits its application on "dilated FCN" [13, 86, 82] and on high resolution datasets like Cityscapes [15]. To address this issue, we propose a more efficient graphic model to achieve faster feature propagation. We find that each DAGs adopted by [66] could be alternatively replaced by two Unidirectional Acyclic Graphs (UAGs), in which the pixels of the same row or column are dealt in parallel with 1D convolution. The proposed UAGs greatly speed up the feature propagation process. Moreover, different from the DAGs that are extremely deep, the proposed UAGs are much shallower and thus alleviate the problem of propagation vanish [57].

Finally, based on the UAG-structured image and the learned boundaries information, we build a boundary-aware feature propagation (BFP) module. In the BFP, local features of the same segment are shared via unimpeded connections to exchange information that achieves feature assimilation, while features of different segments are exchanged under controlled connections with the guidance of learned boundaries. There are several advantages of our proposed boundary-aware feature propagation (BFP) network. First, since the proposed UAGs deal with pixels of the same row or column in parallel, we achieve the propagation process in a high speed. And the UAGs contain much fewer parameters than convolutional methods. Second, as we express boundary detection as classification of a semantic class, lots of parameters and complex module for boundary detection are saved. Third, with the advice of boundary confidence, the local features are propagated in a more motivated way, enhancing the similarity of features belonging to the same segment while keeping the

discrimination of features belonging to different segments.

The main contributions of this paper can be summarized as follows:

- We show that the boundary can be learned as one of semantic categories, which requires little change on network but obtains essential boundary information.
- We propose some unidirectional acyclic graphs (UAGs) to propagate information among high-resolution images with a high speed.
- We propose a boundary-aware feature propagation module to improve the similarity of local features belonging to the same segment region while keeping the discriminative power of features belonging to different segments.
- We achieve new state-of-the-art performance consistently on PASCAL-Context, CamVid, and Cityscapes.

2. Related work

2.1. Scene Segmentation

Scene segmentation (or scene parsing, semantic segmentation) is one of the fundamental problems in computer vision and has drawn lots of attentions. Recently, thanks to the great success of Convolutional Neural Networks (CNNs) in computer vision [42, 68, 71, 52, 25, 72, 27, 80, 26], lots of CNNs based segmentation works have been proposed and have achieved great progress [29, 22, 81, 83, 84, 70, 60]. For example, Long et al. [54] introduce the fully convolutional networks (FCN) in which the fully connected layers in standard CNNs are transformed to convolutional layers. Noh et al. [56] propose deconvolution networks to gradually upsample the coarse features to high resolution. Chen et al. [13] propose to remove some pooling layers (or convolution stride) in CNNs and adopt dilated convolution to retain more spatial information. And some works focus on lightweight network architectures [3, 46] and real-time segmentation [85, 58, 77, 59].

Context aggregation is a hot direction in scene segmentation. For example, Chen et al. [13] propose an atrous spatial pyramid pooling (ASPP) module to aggregate multi-scale context information. Yu et al. [79] employ multiple dilated convolution layers after score maps to exercise multi-scale context aggregation. Zhao et al. [86] introduce pyramid spatial pooling (PSP) to exploit context information from different scale regions. Zhang et al. [82] encode semantic context to network and stress class-dependent feature maps. He et al. [30] propose adaptively pyramid context module to capture global-guided local affinity. Fu et al. [21] integrate local and global dependencies with both spatial and channel attention. Ding et al. [17] employ semantic correlation to infer shape-variant context.

Graphic models have a long history in scene segmentation. Early works construct the graphic model with
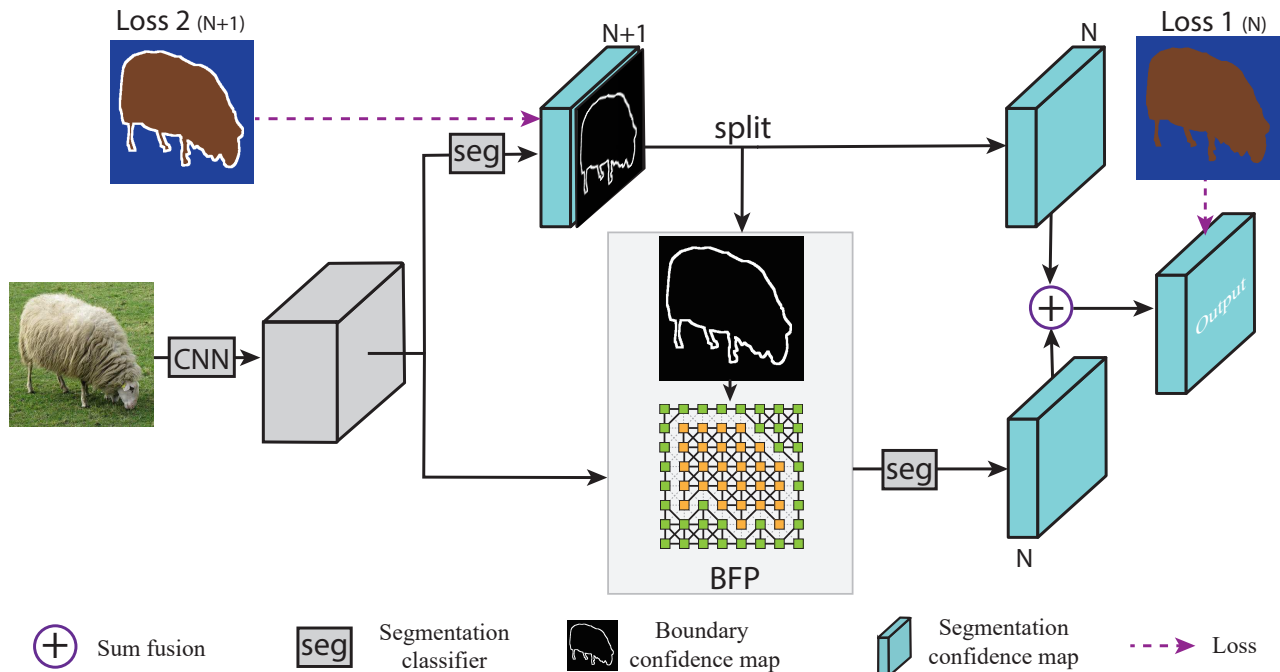
Figure 2. An overview of the proposed approach. We use the ResNet-101 (CNN) with the dilated network strategy [13] as backbone and the proposed boundary-aware feature propagation (BFP) module is placed on the top of CNN. The supervisor of loss 2 is the new ground truth of N+1 classes with an additional boundary class generated from the original ground truth of N classes.

hand-crafted features [24, 51, 75, 69]. Markov Random Fields (MRF) [24, 43, 45] and Conditional Random Fields (CRF) [41, 61, 13, 53] build the dependencies according to the similarities of neighboring pixels. Liang et al. [47] propose to construct graph topology based on superpixel nodes and incorporate long-range context with Graph LSTM. Shuai et al. [66] adopt undirected cyclic graphs (UCGs) to formulate an image and decompose the UCGs with directed acyclic graphs (DAGs). Byeon et al. [9] propose to divide an image into non-overlapping windows and employ a 2d LSTM to construct local and global dependencies. However, most of graph-based method are time-consuming and computationally expensive as they require candidate pre-segments, superpixels, or lots of loops.

In this work, we propose unidirectional acyclic graphs (UAGs), based on which the local features are quickly propagated in parallel. And to construct strong dependencies within the same segment and weak dependencies among different segments, we exploit the learned boundary information to guide the feature propagation within the UAG-structured image.

2.2. Boundary Detection

Boundary detection is a fundamental component for many computer vision tasks and has a long history [1, 19, 40, 35]. For example, Lim et al. [48] propose sketch tokens (ST) and Dollár [20] et al. propose structured edges (SE) based on fast random forests to deal with boundary detec-

tion as local classification problem. Recently the success of CNNs have great improve the performance of boundary detection [5, 6, 33, 64, 74]. Xie et al. [74] employ features from the middle layers of CNNs for boundary detection. Shen et al. [63] propose multi-stage fully convolutional networks for boundary detection of electron microscopy images. These methods target at optimizing the accuracy of boundary detection instead of generating semantic boundary information for high-level tasks. Boundary information could be used for improving segmentation performance. For example, Bertasius et al. [6], Hayder et al. [28], Chen et al. [11] and Kokkinos [38] employ the binary edges to improve the segmentation performance. However, they all employ an additional branch of network for edge detection, which requires more resources and deal with segmentation and boundary detection as two detached tasks. Different from [6, 28, 11], our goal is not to detect the clearly binary boundaries, but to infer a boundary confidence map that represents the probability distributions of high-level boundary layout.

3. Approach

Due to the diverse style and complex layout of scene images, it is necessary to classify every pixel using holistic context information but protect its differentiation from overwhelming by global scene. In this respect, we propose a boundary-aware feature propagation module to arm the

local features with a holistic view of contextual awareness but keep the discriminative power of features for different objects. The overall architecture is shown in Figure 2. We use the dilated FCN (subsampled by 8) based on ResNet-101 [31] as backbone. The supervisor of loss 2 is the boundary-aware ground truth (N+1 classes) that are generated from the original ground truth (N classes).

3.1. Semantic Boundary Detection

Boundary delineation is favourable for meticulous scene parsing. However, because of the various semantic labels and complex layout of objects in segmentation datasets, parsing pixels in the boundary area is always difficult and results in confused prediction. In this work, instead of directly assigning semantic labels to pixels in boundary area, we explore to infer the boundary layout first and improve the segmentation performance with the learned boundary information. Lots of works have contributed to boundary detection [5, 6, 33, 64, 74], but most of them focus on edges that sketch the objects. Different from them, we only focus on the boundaries of semantic objects that are predefined in segmentation datasets. We have observed that boundaries have the property of dramatic changes in RGB and feature information. And the boundary label is easy to be generated from the existing ground truth. Consequently, we assume that the boundary could be viewed as an additional semantic category and simultaneously learned with other existing object categories. As shown in Figure 2, we obtain a new ground truth (Loss 2, N+1 classes) from the original ground truth (Loss 1, N classes) and utilize the new ground truth for supervising the network to learn and infer the boundary layout.

Different from previous boundary detection works that aim to boundary delineation only or deal with segmentation and boundary detection as two detached tasks, our proposed semantic boundary detection is embed with semantic object parsing. Our boundary detection module only targets at boundaries of semantic objects predefined in the training data and generate concise boundary information under interaction with segmentation. These two tasks are combined into one, and they benefit each other. By training them together, the scene segmentation classes help suppress the edges within objects that are not semantic boundary of objects, e.g., edges of eyes. Scene segmentation helps the boundary detection to filter out noise and delineate well-directed boundary, while boundary detection makes the scene segmentation be aware of the important boundary layout information.

3.2. Unidirectional Acyclic Graphs

Context is designed to aggregate wide range of surrounding information, thus it desires a holistic view of the overall image without regard to resolution. One popular way is
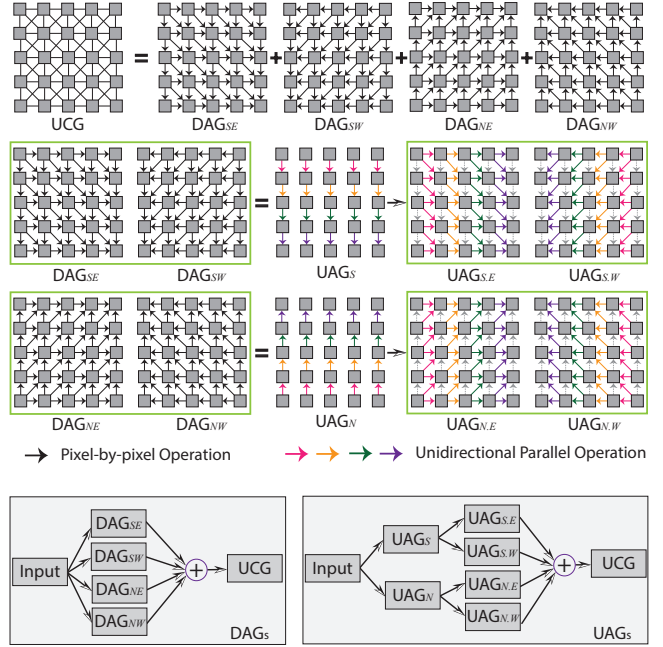

Figure 3. Each point of the DAGs has three different directions. Thus, the DAGs have to scan the image pixel by pixel and consumes lots of time with many loops. We decompose the four DAGs to six Unidirectional Acyclic Graphs (UAGs). Each of UAGs propagates information toward a single direction, which deals with pixels of each row in parallel and then deals with pixels of each column in parallel. For example, the UAG_S is in the south direction only and the $UAG_{S.E}$ is in the east direction only (based on UAG_S).

to employ stacked convolution or dilated convolution to enlarge the receptive field [13, 79, 18], but this consumes large computation resources. The work of [66] proposes DAG-RNN to capture long-range context based on directed acyclic graphs (DAGs). As shown in Figure 3, pixels are locally connected to form a undirected cyclic graph (UCG) to build propagation channels among the whole image. To overcome the loopy property of UCG, the UCG is decomposed to four DAGs with different directions (southeast, southwest, northeast, northwest). However, since each pixel of the DAGs has three different directed connections, the feature propagation based on DAG-structured images has to scan the image pixel by pixel and requires lots of loops. Thus it is very slowly even on low-resolution feature maps, which limits its application on "dilated FCN" [13, 86, 82] and on high resolution datasets like cityscapes [15] and CamVid [8]. To address this issue, we explore to reduce the number of loops and propagate information in parallel. Herein, we propose some Unidirectional Acyclic Graphs (UAGs) as shown in Figure 3, which deal with pixels of each row in parallel and then deals with pixels of each column in parallel. Each DAGs adopted by [66] could
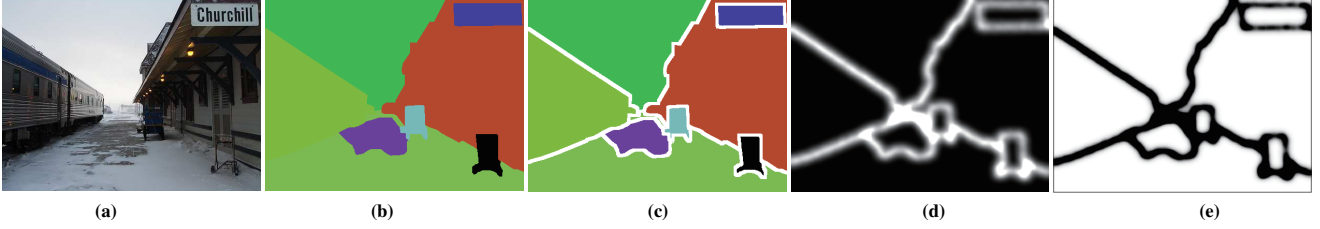
Figure 4. (a) Image; (b) Original ground truth; (c) New generated ground truth: add a boundary class generated from the original ground truth; (d) Boundary confidence map: the probability distribution of the boundary layout; (e) Propagation confidence map: the confidence distribution of propagation.

be alternatively replaced by two UAGs. For example, the DAG$_{SE}$ is decomposed to UAG$_S$ and UAG$_{S.E}$, where UAG$_S$ is south directed that deals with pixels of the same row in parallel and UAG$_{S.E}$ is east directed (after UAG$_S$) that deals with pixels of the same column in parallel. As a result, the number of loops for each DAGs is reduced from $H \times W$ to $H+W$, where H and W are the height and width of feature maps. The proposed UAGs greatly speed up the feature propagation process, which is economic and desired in practice, especially for applications that require high resolution and big eyeshot (e.g., self-driving vehicle). Moreover, due to the pixel-by-pixel operation, the recursive layers in DAGs are very deep that could be unrolled to thousand layers. This causes the problem of propagation vanish [57]. The proposed UAGs are much shallower than DAGs and thus alleviate the the problem of propagation vanish.

3.3. Boundary-Aware Feature Propagation

However, unselective propagation would make the features assimilated, which results in smooth representation and weakens the features' discrimination. To classify features in different objects and stuff in scene segmentation, it is beneficial to improve the feature similarity of the same object while keeping the feature discrimination of different objects. Therefore, we introduce the boundary information into feature propagation to control the information flow between different segments. As shown in Figure 1, with the learned boundary information, we build strong connections for pixels belonging to the same segment but weak connections for different segments. During propagation, more information is passed via strong connections within the same segment and less information flows crossing different segments. In such a way, pixels get more information from other pixels of the same objects and less information from pixels of other objects. Consequently, features of different objects could keep their discrimination while features of the same object trend to be cognate, which is desired for segmentation. The detailed process of proposed boundary-aware feature propagation is presented below.

As our UAGs are unidirectional and in parallel, we formulate the propagation process in 1D here for notation
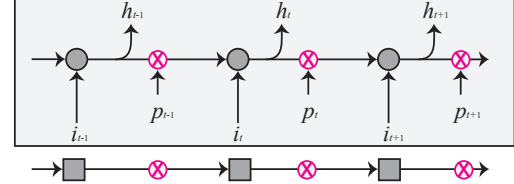

Figure 5. As our UAGs are unidirectional and in parallel, we show the propagation process in 1D here for clarity. i_t represents the feature of pixel at position t, h_t is output (hidden state), p_t is the propagation confidence.

clarity. Extension to 2D/3D is straightforward. We denote the feature of pixel at position t as i_t, and the corresponding output (hidden state) is denoted as h_t. The standard propagation process based on our UAG-structured image is formulated as following:

$$h_t = g(U * i_t + W * h_{t-1} + \delta) \tag{1}$$

where $*$ is 1D convolution operation, U, W are learnable parameters of 1D convolution and δ is learnable bias. g is element-wise nonlinear activation function (we use ReLU).

For the boundary-aware propagation, we first extract the boundary confidence map from the $(N + 1)$ segmentation confidence maps, as shown in Figure 2. We denote the boundary confidence of pixel t as b_t, corresponding to i_t. Based on the boundary confidence map, we generate the propagation confidence map:

$$p_t = 1 - \beta f(\alpha b_t - \gamma) \tag{2}$$

where p_t is the propagation confidence that decides how much the information of pixel t to be passed to next region. α=20 and γ=4 are constant chosen by experience, f is sigmoid function to enhance the boundary, β is a learnable parameter. With the propagation confidence, the propagation process can be reformulated as below:

$$h_t = g(U * i_t + W * h_{t-1}p_{t-1} + \delta) \tag{3}$$

as shown in Figure 5, the propagation is controlled by the boundary and thus models boundary-aware context feature for better parsing of different segments.

For the UAGs that have "two directions", they are also unidirectional and in parallel. For example, $UAG_{S.E}$ is formulated below:

$$h_t^j = g(U * i_t^j + W * h_{t-1}^j p_{t-1}^j + \hat{W} * h_{t-1}^{j-1} p_{t-1}^{j-1} + \delta) \quad (4)$$

There are two hidden states, h_{t-1}^j and h_{t-1}^{j-1}, input to the current cell, where t and j are the denotations of horizontal and vertical axis, respectively. Finally, the hidden states of the corresponding positions in four UAGs (i.e., $UAG_{S.E}$, $UAG_{S.W}$, $UAG_{N.E}$, $UAG_{N.W}$) are fused together to generate the final output, as shown in Figure 3.

One example of boundary confidence map and propagation confidence map is shown in Figure 4. We learn the boundary confidence map under the supervise of new generated ground truth with additional boundary class. To control the progress of feature propagation, propagation confidence map is generated from boundary confidence map. If pixel i_t is in boundary region, then it has a higher boundary probability b_t and hence a smaller propagation probability p_t. Thus, the feature propagation is suppressed and weak signals are passed to next pixel. Otherwise, it has a strong propagation to spread its features to next pixel.

4. Experiments

4.1. Implementation Details

Our Network is implemented based on the public platform Pytorch. We use ResNet-101 [31] with the dilated network strategy [13] (subsampled by 8) as our backbone. In detail, the convolution strides for downsampling in last two blocks are reset to 1 and the convolutions of last two blocks are dilated with dilation rates of 2 and 4 receptively. $Pool$ 5 and layers after it are discarded. The network is trained with mini-batch, batchsize is set to 12 for PASCAL-Context and 8 for Cityscapes and CamVid. Following deeplab-v2 [13], we use the "poly" learning rate policy, where current learning rate Lr_c depends on the base learning rate Lr_b and iterations: $Lr_c = Lr_b \times (1 - \frac{iter}{total\_iter})^{0.9}$. Momentum and weight decay are fixed to 0.9 and 0.0001 respectively. We adopt random horizontal flipping and random resizing between 0.5 and 2 to augment the training data.

Most of the scene segmentation datasets do not provide boundary layout, we use the provided segmentation ground truth to generate boundary-aware ground truth, as shown in Figure 4 (c). As we adopt the dilated FCN as our backbone, the spatial size is downsampled 8 times in encoding process. Thus to avoid the loss of boundary information in feature maps with smallest spatial size, pixels with distance smaller than 9 pixels (i.e., trimap of 18 pixels) to boundary are defined as boundary pixels and their ground truth labels are set to $N+1$, where N is the number of classes in datasets. In our experiments, the over-wide boundary (e.g., trimap of 50

Methods	Input Resolution	# Loops	Time (s)
FCN	480×360	none	0.35
DAGs	480×360	10800	17.92
UAGs	480×360	300	0.47
FCN	960×720	none	0.42
DAGs	960×720	43200	56.97
UAGs	960×720	600	0.76

Table 1. Inference speed comparison of FCN (baseline), DAGs, and UAGs on dilated ResNet-101 with different resolution inputs.

pixels) squeezes small objects and weakens the function of boundary in feature propagation. We evaluate our network with mean Intersection-over-Union (**mIoU**). Mathematical definition of mIoU please refer to [54].

4.2. Efficiency Analysis

To evaluate the speed of the proposed UAGs, we report in Table 1 the inference time of UAGs and compared it with DAGs [66] on different resolution of input images, based on dilated FCN (subsampled by 8). The number of loops is also recorded. Different from DAGs that have to scan the image pixel by pixel, the proposed UAGs deal with pixels of each row/column in parallel, hence they save a lot of loops. As shown in Table 1, the UAGs contain much less loops than DAGs and consequently they are much faster than DAGs. Especially with high resolution (e.g., 960×720), DAGs are very slow and time-consuming. The speed of DAGs is highly related with the input resolution that determines the number of loops, thus DAGs are not suitable of high-resolution datasets (e.g., Cityscapes [15]) and FCN with dilated network strategy [13]. Besides the inference time, training of DAGs based on dilated FCN requires one hundred times more GPU hours than our proposed UAGs, which also shows the high efficiency of our approach. To quantitatively compare the segmentation performances of DAGs and proposed UAGs, we evaluate them on VGG-16 [67] with encoder-decoder strategy, exactly in the same way like that in [66]. DAGs and UAGs achieve almost the same results on PASCAL-Context (UAGs 43.0% vs. DAGs 42.6%). This shows that the proposed UAGs realize the same function as DAGs but with a much faster speed.

4.3. Ablation Studies

We show the detailed results of ablation studies of the proposed approach in Table 2. The proposed UAGs harvests local features and propagate them throughout the whole image to achieve a holistic context aggregation, which greatly improves the segmentation performance from the baseline (dilated FCN). Then, based on the UAGs, we learn the boundary information and inject it to the propagation process to control the information flow between different regions. With the boundary information, the UAGs build stronger connections of pixels within the same segment
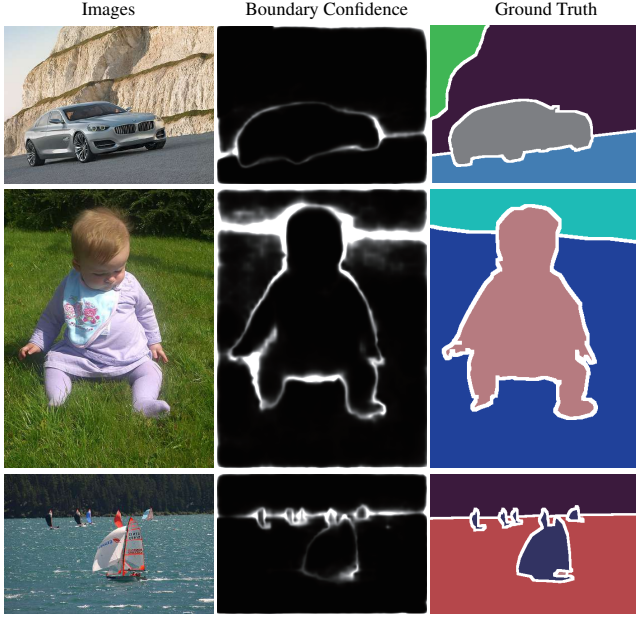
| Images | Boundary Confidence | Ground Truth |

Figure 6. Qualitative examples of inferred boundary map.

Figure 7. Segmentation performance within band (trimap) around boundaries.

Methods	Backbone	UAGs	Boundary	MS	mIoU
FCN	ResNet-50				41.2
BFP	ResNet-50	✓			49.8
BFP	ResNet-101	✓			50.8
BFP	ResNet-101	✓	✓		52.8
BFP	ResNet-101	✓	✓	✓	53.6

Table 2. Ablation Study of Boundary-aware Feature Propagation (BFP) Network on PASCAL-Context. Baseline is dilated FCN and MS means multi-scale testing.

Methods	mIoU
O2P [10]	18.1
FCN-8s [62]	39.1
BoxSup [16]	40.5
HO-CRF [2]	41.3
PixelNet [4]	41.4
DAG-RNN [66]	43.7
EFCN [65]	45.0
DeepLab-v2+CRF [13]	45.7
RefineNet [50]	47.3
MSCI [49]	50.3
CCL+GMA [18]	51.6
EncNet [82]	51.7
BFP (ours)	**53.6**

Table 3. Testing accuracies on PASCAL-Context.

Methods	mIoU
DeconvNet [56]	48.9
SegNet [3]	50.2
FCN-8s [54]	52.0
DeepLab [12]	54.7
DilatedNet [79]	65.3
Dilation+FSO [44]	66.1
G-FRNet [34]	68.0
Dense-Decoder [7]	70.9
BFP (ours)	**74.1**

Table 4. Testing accuracies on CamVid.

and weaker connections between different segments. Thus, features of the same segment become more similar but features of different segments remain discriminative.

We also visualize some examples of the inferred boundary confidence map in Figure 6. As shown in Figure 6, the inferred boundary map mainly involves boundaries between the semantic segments predefined in datasets instead of other edge information. Therefore, it contains the desired boundary layout of semantic objects and could be used for control of the feature propagation throughout image. The results in Table 2 show the effectiveness of the boundary-aware feature propagation (BFP) network. Following [37, 11], we evaluate the performance of BFP near boundaries, as shown in Figure 7. We compute the mIoU for the regions within different bands of boundaries.

DT-EdgeNet [11] is the most related to BFP. However, BFP learns the boundary as one of semantic classes and is trained end-to-end, while DT-EdgeNet learns edge with additional EdgeNet and requires a two-step training process for DeepLab and EdgeNet. As shown in Figure 6, Our learned boundaries response less to object interior edges than DT-EdgeNet. BFP is proposed to perform feature propagation that is some kind of contextual feature modeling, while DT is used to refine the segmentation scores. We use the DT to filter the segmentation scores of BFP and this brings 0.7% performance gain, which shows that the DT and BFP are complementary.

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

PASCAL-Context [55] provides pixel-wise segmentation annotation for the whole scene image. There are 4998 training images a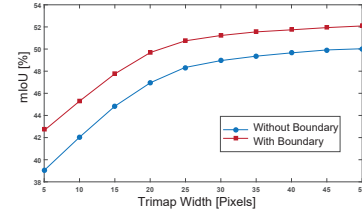nd 5105 testing images in PASCAL-Context. Following [55], we use the most common 59 classes for evaluation. Testing accuracies of PASCAL-Context are shown in Table 3, which shows that the proposed BFP outperforms the state-of-the-art works by a large margin.

CamVid [8] is a road scene image segmentation dataset which provides dense pixel-wise annotations for 11 semantic categories. There are 367 training images, 101 validation images and 233 testing images. The testing results are

Methods	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
FCN [62]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
DPN [53]	97.5	78.5	89.5	40.4	45.9	51.1	56.8	65.3	91.5	69.4	94.5	77.5	54.2	92.5	44.5	53.4	49.9	52.1	64.8	66.8
LRR [23]	97.7	79.9	90.7	44.4	48.6	58.6	68.2	72.0	92.5	69.3	94.7	81.6	60.0	94.0	43.6	56.8	47.2	54.8	69.7	69.7
Deeplabv2 [13]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
RefineNet [50]	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70.0	73.6
DepthSet [39]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	78.2
ResNet-38 [73]	98.5	85.7	93.1	55.5	59.1	67.1	74.8	78.7	93.7	72.6	95.5	86.6	69.2	95.7	64.5	78.8	74.1	69.0	76.7	78.4
PSPNet [86]	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
AAF [36]	98.5	85.6	93.0	53.8	59.0	65.9	75.0	78.4	93.7	72.4	**95.6**	86.4	70.5	95.9	73.9	82.7	76.9	68.7	76.4	79.1
DFN [78]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	79.3
PSANet [87]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	80.1
DenseASPP [76]	98.7	**87.1**	93.4	**60.7**	62.7	65.6	74.6	78.5	93.6	72.5	95.4	86.2	71.9	96.0	**78.0**	**90.3**	80.7	**69.7**	76.8	80.6
BFP (ours)	**98.7**	87.0	**93.5**	59.8	**63.4**	**68.9**	76.8	80.9	93.7	72.8	95.5	**87.0**	72.1	**96.0**	77.6	89.0	**86.9**	69.2	**77.6**	**81.4**

Table 5. Category-wise performance comparison on the Cityscapes test set. Note that the DenseAspp [76] uses stronger backbone DenseNet-161 [32] than Resnet-101 [31] we adopt as backbone.

shown in Table 4. It shows again that the proposed BFP outperforms previous state-of-the-arts by a large margin.

Cityscapes [15] is a recent street scene dataset which contains 5000 high-resolution (1024×2048) images with pixel-level fine annotations. There are 2975 training images, 500 validation images and 1525 testing images. 19 classes (e.g., roads, bicycles and cars) are considered for evaluation on the testing sever provided by the organizers. The Category-wise results are shown in Table 5. Our BFP is only trained on fine annotations, while [14] also uses coarse annotations for training. Some segmentation examples on Cityscapes are shown in Figure 8.

5. Conclusion

In this work, we address the challenging issue of scene segmentation. In order to improve the feature similarity of the same segment while keeping the feature discrim-ination of different segments, we explore to propagate features throughout the image under the control of inferred boundaries. Towards to this, we first propose to learn the boundary as an additional semantic class to enable the network to be aware of the boundary layout. Then, in order to structurize the image to define the order rules for feature propagation, we propose some unidirectional acyclic graphs (UAGs) to model the function of undirected cyclic graphs (UCGs) in a much more efficient way than DAGs. Based on the proposed UAGs, holistic context is aggregated via harvesting and propagating the local features throughout the whole image efficiently. Finally, we propose a boundary-aware feature propagation (BFP) network to detect and utilize the boundary information for controlling the feature propagation of the UAG-structured image. The proposed BFP is capable of improving the similarity of local features belonging to the same segment region while keeping the discriminative power of features belonging to
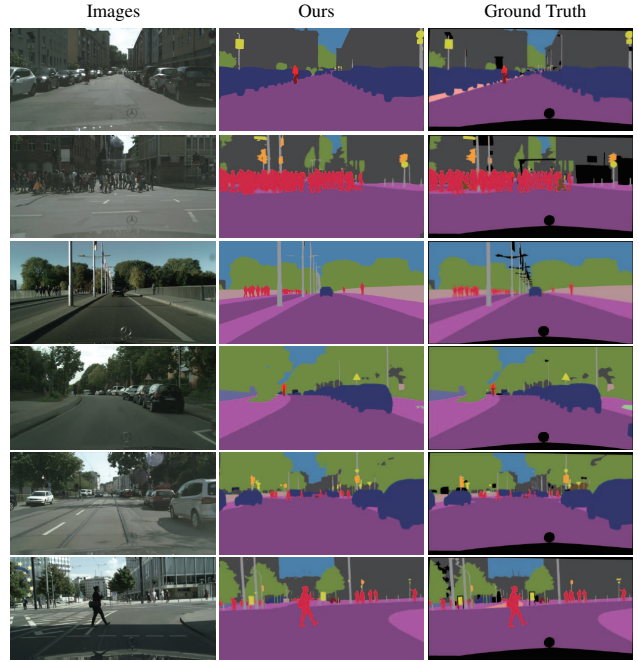

Figure 8. Qualitative segmentation examples on Cityscapes.

different segments. We evaluate the proposed boundary-aware feature propagation network on three changeling se-mantic segmentation datasets, PASCAL-Context, CamVid, and Cityscapes, which show that the proposed BFP achieves new state-of-the-art segmentation performance consistently.

References

[1] 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, 2011.

[2] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*. Springer, 2016.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017.

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

[5] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 4380–4389, 2015.

[6] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 504–512, 2015.

[7] Piotr Bilinski and Victor Prisacariu. Dense decoder shortcut connections for single-pass semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, 2008.

[9] Wonmin Byeon, Thomas M. Breuel, Federico Raue, and Marcus Liwicki. Scene labeling with lstm recurrent neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. *Computer Vision–ECCV 2012*, 2012.

[11] 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 *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, 2016.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[13] 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.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018.

[15] 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*, 2016.

[16] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[17] 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 (CVPR)*, pages 8885–8894, June 2019.

[18] 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 (CVPR)*, pages 2393–2402, June 2018.

[19] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1848, 2013.

[20] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1558–1570, 2015.

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

[22] 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*, 2019.

[23] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*. Springer, 2016.

[24] 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.

[25] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *ECCV*, 2018.

[26] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *ICCV*, 2019.

[27] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019.

[28] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5704, 2017.

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

[30] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[32] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] Jyh-Jing Hwang and Tyng-Luh Liu. Pixel-wise deep learning for contour detection. *arXiv:1504.01989*, 2015.

[34] Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Crisp boundary detection using pointwise mutual information. In *European Conference on Computer Vision*, pages 799–814. Springer, 2014.

[36] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X. Yu. Adaptive affinity fields for semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[37] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[38] Iasonas Kokkinos. Pushing the boundaries of boundary detection using deep learning. *arXiv:1511.07386*, 2015.

[39] Shu Kong and Charless C Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[40] Scott Konishi, Alan L. Yuille, James M. Coughlan, and Song Chun Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):57–74, 2003.

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

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

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

[44] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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

[46] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3202, 2017.

[47] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[48] Joseph J Lim, C Lawrence Zitnick, and Piotr Dollár. Sketch tokens: A learned mid-level representation for contour and object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3158–3165, 2013.

[49] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[50] 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 (CVPR)*, 2017.

[51] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5), 2011.

[52] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex Kot Chichung. Feature boosting network for 3d pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[53] 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*, 2015.

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

[55] 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 (CVPR)*, 2014.

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

[57] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

[58] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv:1606.02147*, 2016.

[59] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2018.

[60] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 5639–5647, 2018.

[61] Chris Russell, Pushmeet Kohli, Philip HS Torr, et al. Associative hierarchical crfs for object class image segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 739–746. IEEE, 2009.

[62] Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[63] Wei Shen, Bin Wang, Yuan Jiang, Yan Wang, and Alan Yuille. Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2391–2400, 2017.

[64] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2015.

[65] 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, 2019.

[66] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1480–1493, 2018.

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

[68] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[69] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[70] Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In *Proceedings of the International Conference on Computer Vision*, 2019.

[71] Tiantian Wang, Lihe Zhang, Huchuan Lu, Chong Sun, and Jinqing Qi. Kernelized subspace ranking for saliency detection. In *ECCV*, pages 450–466, 2016.

[72] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3135, 2018.

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

[74] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[75] Jimei Yang, Brian Price, Scott Cohen, and Ming-Hsuan Yang. Context driven scene parsing with attention to rare classes. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[76] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[77] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, 2018.

[78] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *arXiv:1804.09337*, 2018.

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

[80] Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, and Ali Borji. Learning to promote saliency detectors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[81] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the International Conference on Computer Vision*, 2019.

[82] 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 (CVPR)*, 2018.

[83] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[84] Lu Zhang, Zhe Lin, Jianming Zhang, Huchuan Lu, and You He. Fast video object segmentation via dynamic targeting network. In *Proceedings of the International Conference on Computer Vision*, 2019.

[85] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[86] 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 (CVPR)*, 2017.

[87] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Pointwise spatial attention network for scene parsing. In *ECCV*, 2018.