Toward Achieving Robust Low-Level and High-Level Scene Parsing

Bing Shuai, Henghui Ding, Ting Liu, Gang Wang, *Senior Member, IEEE*, and Xudong Jiang, *Senior Member, IEEE*

***Abstract*— In this paper, we address the challenging task of scene segmentation. We first discuss and compare two widely used approaches to retain detailed spatial information from pre-trained convolutional context network (CNN)—“dilation” and “skip”. Then, we demonstrate that the parsing performance of “skip” network can be noticeably improved by modifying the parameterization of skip layers. Furthermore, we introduce a “dense skip” architecture to retain a rich set of low-level information from the pre-trained CNN, which is essential to improve the low-level parsing performance. Meanwhile, we propose a CCN and place it on top of pre-trained CNNs, which is used to aggregate contexts for high-level feature maps so that robust high-level parsing can be achieved. We name our segmentation network enhanced fully convolutional network (EFCN) based on its significantly enhanced structure over FCN. Extensive experimental studies justify each contribution separately. Without bells and whistles, EFCN achieves state-of-the-arts on segmentation datasets of ADE20K, Pascal Context, SUN-RGBD, and Pascal VOC 2012.**

***Index Terms*— Scene parsing, convolution neural network, convolutional context network, fully convolutional network, skip layers.**

I. Introduction

SCENE segmentation refers to parsing a scene image into a set of coherent semantic regions. It is a challenging task that implicitly subsumes object recognition as well as boundary delineation. It demands multi-level parsing ranging from low-level (e.g., boundary localization) to high-level (e.g., object recognition). Thus, it’s understandable that a well-performed segmentation network should effectively incorporate different scale information.

State-of-the-art segmentation networks are based on pre-trained classification network (e.g., CNN). However, detailed spatial information are largely lost in its output feature maps. In order to incorporate the detailed low-level features, researchers usually adapt the architecture of pre-trained CNN [4], [9], [21], [37] based on two approaches: (1) “dilation” - it removes `pool` layers (in pre-trained CNN) and then performs subsequent convolution operations with higher dilation factors (i.e. stride rate); or (2) “skip” - it adds skip branches (usually linear classifier) from early layers of pre-trained CNN such that low-level features are explicitly incorporated into image parsing. Recent literatures [4], [34], [37] as well as our experiments suggest that “dilation” network[1] outperforms its “skip” counterpart. However, there are some defects using the dilation to retain spatial information. First, the resolution of feature maps does not decrease after the layer where dilated (atrous) convolution is applied, so it requires more computations to process “dilation” network. Similarly, memory consumption is also much more severe in “dilation” network. These motivate us to explore using skip layers to retain spatial information. In this paper, we demonstrate that the parsing performance of “skip” network can be significantly improved by re-parameterizing the skip layers. Importantly, skip layers entail very few computation overhead. Hence, “skip” segmentation network can be further combined with a complex up sampling module to recover detailed information. In contrast, “dilation” networks [4] usually exclude extra up sampling module due to efficiency issues. Next, we introduce a “dense skip” architecture to retain rich set of low-level information from pre-trained CNN. In detail, we consider every possible informative lower-level feature maps into semantic parsing. These dense skip layers enable the segmentation network to incorporate very rich low-level contexts, which is essential to improve its performance on low-level parsing. In comparison with the popular “dilation” network, we demonstrate that the “dense skip” architecture is more effective as well as more efficient in terms of retaining low-level detailed information from pre-trained CNN. Besides, dense skip layers offer an effective approach to address the large scale variation of objects in scene segmentation.

Feature maps generated by pre-trained CNN are usually robust enough for image-level visual recognition, but they are not equally discriminative and representative at every spatial location. Take images from Fig. 1 for example, local features for upper-right “train” regions (first image) and lower-right “ground” regions (third image) are not discriminative such

Manuscript received July 14, 2017; revised September 13, 2018 and February 20, 2018; accepted January 10, 2018. Date of publication October 31, 2018; date of current version November 21, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gang Hua. *(Bing Shuai and Henghui Ding are co-first authors.) (Corresponding author: Xudong Jiang.)*

B. Shuai was with Nanyang Technological University, Singapore 639798. He is now with Amazon, Seattle, WA 98101 USA (e-mail: bshuai001@ntu.edu.sg).

H. Ding and X. Jiang are with the Rapid-Rich Object Search Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ding0093@ntu.edu.sg; exdjiang@ntu.edu.sg).

T. Liu and G. Wang are with Alibaba AI Labs, Hangzhou 311121, China (e-mail: brooks.lt@alibaba-inc.com; wg134231@alibaba-inc.com).

Digital Object Identifier 10.1109/TIP.2018.2878975

[1] Hereafter, we simply use “skip” and “dilation” network to notate the above two network architectures.

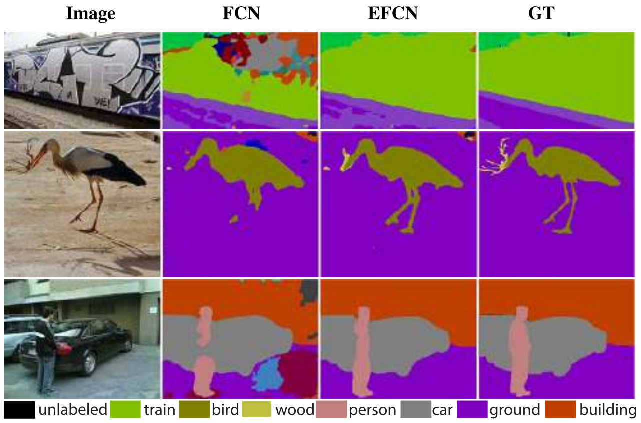


Fig. 1. (Best viewed in color) Semantic segmentation demands robust high-level as well as low-level parsing. EFCN outperforms FCN for both the high-level smoothing/recognition ("train" and "ground" in the first and third example) and the low-level boundary localization (e.g., bird legs in second example).

that their unary predictions are noisy. Thus, feature context aggregation (`CA`) is essential to improve their representation capability. Specifically, context aggregation allows local features to embed their neighborhood informative context so that even a linear classifier (`fc` layer) can easily distinguish their genuine categories. In this paper, we propose a convolutional context network (CCN) and place it on top of pre-trained CNN to aggregate context for high-level feature maps. CCN is employed to contextualize high-level features using basic convolutional kernels to achieve high-level consistency. Different from atrous convolution [4], [37], there will be no "holes" during the convolution operations. Thus, it can incorporate every feature within the contextual window into corresponding features. CCN could also aggregate multi-scale context information to overcome the challenge of the existence of objects at multiple scales. It is empirically demonstrated that CCN is able to significantly improve the parsing performance of segmentation networks. Meanwhile, CCN delivers a more effective `CA` module than many state-of-the-arts [4], [19], [28], [37] including Atrous Spatial Pyramid Pooling (ASPP) in DeepLab [4] and DAG-RNN [28].

Our segmentation network is fully convolutional, and it has dense skip layers from pre-trained CNN and CCN. Thus, we name it enhanced FCN (EFCN) based on its enhanced structure over FCN. The qualitative parsing results in Fig. 1 illustrate that EFCN outperforms FCN in terms of both high-level as well as low-level parsing. Overall, EFCN presents a strong segmentation network architecture, which achieves state-of-the-art performance on standard semantic segmentation datasets including ADE20K [40], Pascal Context [23], SUN-RGBD [32] and VOC 2012 [6]. In summary, this paper makes the following contributions to yield the network architecture of EFCN.

- We study and compare two adaptation approaches to retain low-level information from pre-trained CNN - "dilation" and "skip". By modifying the parameterization of skip layers, we improve the parsing performance of "skip" network significantly.
- We further propose a "dense skip" architecture to retain rich set of low-level information from pre-trained CNN, which is demonstrated to be more effective as well as more efficient than its "dilation" counterpart.
- We propose a convolutional context network (CCN) to aggregate context for high-level feature maps. We show that CCN is very effective to boost the parsing performance of segmentation networks.

Note that these contributions can also be integrated to other segmentation network architectures, such as SegNet [1], PSPNet [38] and DeconvNet [24]. We also compare qualitative results of EFCN with state-of-the-art segmentation networks, which gives deeper insight to understand our contributions. We will release the training code and the trained model upon the acceptance of this paper.

II. Related Work

A. Retaining Detailed Information

State-of-the-art segmentation networks [4], [5], [21], [35], [38] are adapted from pre-trained classification networks (e.g., CNN trained from ImageNet [26]). Thus, it's essential to address the issue that the detailed low-level information is progressively lost in higher-level feature maps. The seminal work - FCN [21] - demonstrated that more detailed parsing maps can be generated if lower-level feature maps are incorporated into predictions. From the perspective of architecture designs, this is achieved by adding skip layers from lower-level feature maps of pre-trained CNN. Recently, DeepLab [4] and DilatedNet [37] advocated to mitigate the loss of detailed spatial information by removing pooling layers in pre-trained CNN. Afterwards, they perform the subsequent convolution operations with higher dilation factors (i.e. stride rate) so that parameters in pre-trained CNN can be reusable. Most recent segmentation networks are based on the latter architecture design due to its superior parsing performance, such as [4], [34], [35], and [38]. In this paper, we demonstrate that the performance of "skip" network can be significantly enhanced by re-parameterizing skip layers. Moreover, we propose "dense skip" architecture to retain rich set of low-level detailed information, which is shown to be more effective as well as more efficient than its "dilation" counterpart.

In addition, researchers have also been dedicated to developing networks (i.e. decoder network) [1], [24], [34] to recover the lost spatial details. As it will be shown later, our work in this paper is orthogonal to this line of research.

B. Contextual Modeling

One branch of work introduces new computational layers to achieve contextual modelling, which are usually placed on top of pre-trained CNN to enhance high-level parsing performance. For example, Liu *et al.* [20] adopted local convolution layers to approximate the mean field algorithm for pairwise terms in deep parsing network (DPN). Lin *et al.* [17] inserted convolution layers to model the semantic compatibility between image regions. Visin *et al.* [33] and Shuai *et al.* [28] employed recurrent neural networks (RNNs) to propagate local context in feature maps. Chen *et al.* [4] proposed an atrous spatial pyramid pooling (ASPP) network to aggregate multi-scale context for feature maps. Recently, Yu and Koltun [37] introduced a convolutional network using dilated convolution kernels to perform context aggregation over class likelihood maps. All these computational layers are designed to encode extra context into local features so that their representation capabilities are enhanced. In this paper, we propose convolutional context network (CCN) to achieve this functionality, and we demonstrate that CCN is more effective than many state-of-the-art CA modules. An in-depth comparison between CCN and its counterparts is elaborated in Section III-C.

Different from [17], [20], and [37], our context network has multiple shortcut connections, which allow it to have deeper architecture by mitigating its optimization difficulties. Thus, our context network is able to expand the receptive fields significantly larger than those networks. Moreover, shortcut connections enable EFCN to fuse rich-scale contextual predictions, whereas the network architectures in [17], [20], and [37] don't have such property. In contrast to [28], EFCN is more efficient as features (in feature maps) are processed in parallel rather than sequentially as in [28].

Another representative branch of work leverages fully connected CRF [15] (CRF) to contextualize the unary predictions of segmentation networks. For example, Chen *et al.* [4] applied CRF to the unary predictions of DeepLab network, and they observed obvious improved visual quality of parsing maps. Subsequently, Zheng *et al.* [39] formulated CRF as a Recurrent Neural Network (CRF-RNN) so that it can be jointly trained with segmentation networks. Even though CRF is effective towards refining the label maps, they are more like a post-processing refinement step. As demonstrated in [4] and [37], these works are expected to be orthogonal to the contribution of the proposed CCN.

C. Multi-Scale Aggregation

Multi-scale aggregation (feature or class-likelihood) is essential to address the large scale/size variation of objects in semantic segmentation. Technically, there are multiple ways to achieve this goal. For example, the multi-resolution approach [7], [17] generates an image pyramid and then concatenates the corresponding features from different resolutions. Alternatively, the hyper-column approach [2], [11] combines different-levels of convolutional feature maps. Both approaches are expected to incur either higher computation time or larger memory consumption. Thus, in practice, they are limited to aggregate only few scale contextual predictions. The skip layers introduced in [21] locally classify feature maps of different scales and then fuse their predictions, which is proven effective as well as economic to achieve multi-scale aggregation. In this paper, we introduce dense skip approach to enable segmentation networks to fuse rich-scale contextual predictions, which is critical to deliver detailed parsing maps. As linear fusion strategy is adopted in FCN, more advanced fusion method like in [8] can be utilized to further improve the parsing performance.

III. Segmentation Networks

Scene segmentation requires both robust low-level and high-level parsing. Specifically, low-level spatial information is important to detect small-size objects as well as to delineate their boundaries. On the other hand, context-aware and high-level feature representation is essential to recognize stuff regions (e.g., 'building' and 'ground' in Fig. 1) and appearance-inconsistent or zoom-in objects (e.g., 'train' in Fig. 1). In this paper, we discuss how to integrate robust high-level and low-level parsing in an efficient segmentation network.

A. Two Approaches to Retain Detailed Spatial Information: Dilation and Skip

Pre-trained CNN (i.e. local representation module) usually outputs high-level semantic features that are essential for robust object recognition. However, low-level and mid-level features are also important for image parsing. For example, localizing small-size objects and delineating object boundaries requires robust low-level visual recognition. Thus, how to retain and incorporate the detailed spatial information from pre-trained CNN remains a promising research direction. Overall, two prominent approaches arise.

- **Dilation**. This line of work is represented by DeepLab [4] and DilatedNet [37]. These two networks remove `pool` layers, and modify subsequent `conv` kernels with dilated (atrous) `conv`. By doing so, resolution of feature maps doesn't reduce significantly, and hence the detailed spatial information is preserved.
- **Skip**. The seminal work FCN [21] adopts this architecture design. Specifically, skip layers are used to fuse predictions from early feature maps of pre-trained CNN. Thus, lower-level features are explicitly incorporated in semantic segmentation.

We experimentally compare and discuss the two approaches on ADE20K [40]. Results are summarized in Table I. Consistent with current findings in [4] and [37], our experiments show that "dilation" network shows superior parsing performance than its "skip" counterpart. However, the inference speed of "dilation" networks is significantly slower, as they need to process denser feature maps.

To compare these two approaches, we modify the pre-trained CNN in the corresponding way as suggested by these methods. Specifically for VGG-16, we remove `pool4` and `pool5`, and the dilation factors for `conv` kernels after these layers are set to 2 and 4 respectively. Alternatively, we add two skip branches emanating from `pool3` and `pool4`. All of the other modules in the segmentation networks are fixed identically, and we have compared these modifications for two segmentation networks: FCN and DeepLab. The results on ADE20K dataset are summarized in Table I.

As shown in Table I, the two modifications both perform well for these networks, which demonstrates that both approaches can retain some useful information to improve parsing performance. However, it's important to mention that "dilation" network is significantly slower than "skip" network. Considering that the resolution of feature maps doesn't

TABLE I

TWO APPROACHES - "DILATION" AND "SKIP" - ARE USED TO ADAPT PRE-TRAINED CNN (IN THIS CASE, VGG-16 [31]). RESULTS OF THEIR CORRESPONDING NETWORKS ARE EVALUATED ON ADE20K DATASET [40]. NETWORKS ARE TRAINED UNDER THE SAME CONDITION (DETAILS ARE ELABORATED IN SECTION IV) EXCEPT THAT FCN-8S (ORIG) IS TRAINED WITHOUT CLASS-WEIGHTED LOSS [28]. SPEED REFERS TO THE INFERENCE TIME OF SEGMENTATION NETWORKS ON A SINGLE NVIDIA TITAN X

Networks	Type	IOU	Speed (hz)
FCN-8s (Orig) [40]	skip	29.5%	**10.0**
FCN-8s	skip	31.4%	**10.0**
FCN	dilation	**32.8**%	3.60
FCN-8s	our skip	32.6%	**10.0**
DeepLab-v2	skip	33.5%	**10.8**
DeepLab-v2	dilation	**35.1**%	7.30
DeepLab-v2	our skip	34.7%	**10.8**

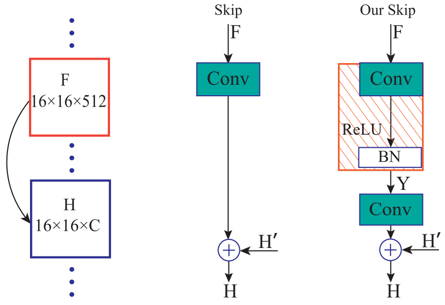


Fig. 2. Skip layers of segmentation networks. The right figure depicts the detailed computational blocks of two parameterizations of skip layers. Mathematically, $\mathbb{H} = \mathbb{H}' + S(\mathbb{F}, \Theta)$, where $\mathbb{F}$ and $\mathbb{H}$ denotes input feature map (with the dimensionality of *height* × *width* × *#channels*) and its class-likelihood map respectively, $S(\cdot, \cdot)$ is the computational function of skip layer and Θ is their learnable parameters. $\mathbb{H}'$ represents class-likelihood maps generated from other branches.

decrease after `pool3` layer, it requires significantly more computations. In accordance, memory consumption is also more severe in "dilation" network.

This result motivates us explore designing efficient "skip" network architecture to enhance its capability in retaining detailed spatial information.

We found that the inferior results of "skip" networks are partly caused by the less-well trained skip layers. Referring to Fig. 2, the gradients that are back-propagated to pre-trained CNN are $\Delta\mathbb{F} = \frac{\partial \mathcal{L}}{\partial \mathbb{H}} \frac{\partial \mathbb{H}}{\partial \mathbb{F}}$, where $\mathcal{L}$ refers to the class-weighted loss [28]. Consider the usual parameterization of skip layer (1×1 `conv` kernel - parameterized by W), $\Delta\mathbb{F} = W^T \frac{\partial \mathcal{L}}{\partial \mathbb{H}}$, the error signals $\frac{\partial \mathcal{L}}{\partial \mathbb{H}}$ vary in a large range and strong error signals are not functionally modulated before they are propagated to the early layers of the pre-trained CNN. Thus, gradient norm $||\Delta\mathbb{F}||$ of large error signals can be very large, which fluctuates the network training. In practice, a simple magnitude smaller learning rate is usually used for skip layers [27] to prevent the network training from divergence. However, this suppresses some small error signals $\frac{\partial \mathcal{L}}{\partial \mathbb{H}}$ in the training, which results in a less-well trained skip layers in segmentation networks.

TABLE II

PARSING PERFORMANCE COMPARISON BETWEEN "SPARSE SKIP" AND "DENSE SKIP" NETWORK ARCHITECTURES ON ADE20K DATASET [40]. HERE OUR NEW PARAMETERIZED SKIP LAYERS ARE APPLIED TO BOTH SPARSE AND DENSE SKIPS

Networks	Type	IOU	Speed
FCN	sparse skip	32.6%	10.0
FCN	dense skip	**33.6**%	9.10
DeepLab-v2	sparse skip	34.7%	10.8
DeepLab-v2	dense skip	**35.4**%	10.1

To address this issue, we use a 2-layer convolutional network with batch normalization (`BN`) [13] to parameterize skip layers (i.e. `Conv+BN+ReLU+Conv`). Specifically, batch normalization is used to stabilize the back-propagated error signals $||\Delta\mathbb{F}||$. Now the gradients that back propagated to pre-trained CNN are $\Delta\mathbb{F} = \frac{\partial \mathcal{L}}{\partial \mathbb{H}} \frac{\partial \mathbb{H}}{\partial \mathbb{F}} = W^T \frac{\partial \mathcal{L}}{\partial \mathbb{H}} \frac{\partial \mathbb{Y}}{\partial \mathbb{F}}$, where $\mathbb{Y}$ refers to the output of the newly added layers `Conv+BN+ReLU` and W is the parameter of the last 1×1 classifier `Conv` layer. $\frac{\partial \mathbb{Y}}{\partial \mathbb{F}}$ plays a modulative role to the error signals $\frac{\partial \mathcal{L}}{\partial \mathbb{H}}$ and we can view it as an adjustable learning rate, which is adaptive to the error signal. Thus, the skip layers can be better trained without posing optimization issues.

The role of our newly added layers is tested in the two well-known segmentation networks (FCN [27] and DeepLab-v2 [4]).[2] Their results are shown in Table I. Performance of the same "skip" networks are significantly boosted (> 1% IOU) after our new parameterized skip layers are used. As our parameterization of skip layers bring negligible extra computations, the inference of "our skip" networks remains efficient. Importantly, "our skip" networks achieve comparable parsing performance with its "dilation" counterpart.

B. Dense Skip: A More Effective Approach to Retain Spatial Information

In order to retain rich set of low-level information from pre-trained CNN, we propose "dense skip" architecture. In this sense, the corresponding network is expected to perform better on low-level parsing. Besides, it is able to incorporate very rich-scale context to make semantic predictions, which is essential to address the large scale variation of objects in scene segmentation. Here, the "dense skip" network adds skip layers for each intermediate feature map after `pool3` so that it engages the same scale information as the "dilation" network. Accordingly, we adapt two segmentation networks (FCN [27] and DeepLab-v2 [4]) to "dense skip" architecture. Note that original FCN is a "sparse skip" network. We compare their performance in Table II. As expected, the "dense skip" architecture outperforms the conventional "sparse skip" counterpart by a noticeable margin, and its inference speed remains competitively fast. In comparison with "dilation" network

[2] The architecture of DeepLab-v2 [4] is illustrated in Fig. 5. In order to adapt its "dilation" architecture to "skip", we revise the dilation factors (stride rate) in the four branches of ASPP to {1, 3, 4, 6} respectively. Their corresponding dilation factors in "dilation" network are {4, 12, 16, 24}, which are close to their original settings {6, 12, 18, 24} in [4].

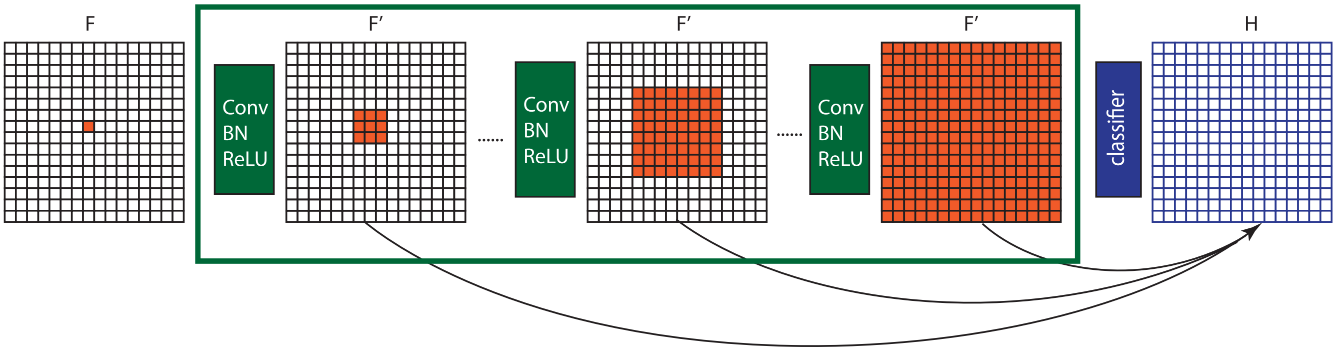


Fig. 3. **Convolutional context network** (CCN) is a convolutional network with dense skip layers. Contextual field (orange coverage in $\mathbb{F}'$) of feature maps are progressively expanded. Note that features in $\mathbb{F}'$ incorporate all elements in $\mathbb{F}$.

(c.f. Table I), the proposed "dense skip" architecture offers a more effective and more efficient framework to preserve detailed spatial information.

If the efficiency is not an issue, it's technically feasible to integrate "dilation" with "dense skip" architecture to a unified network so that lower-level information can be retained from different and possibly complementary manners. We leave this architecture exploration for future work.

C. Context Aggregation for High-Level Features

We have discussed how to adapt pre-trained CNN to retain the detailed spatial information. Meanwhile, robust high-level visual recognition is essential to achieve good parsing performance. Feature maps generated by pre-trained CNN generally encode high-level semantic information, which are globally discriminative for object/scene recognition [12], [31]. However, they are not equally discriminative at every spatial location, especially for zoom-in objects and in "stuff" regions. Take the first example in Fig. 1 as illustration, the "train" pixels are visually inconsistent, and they can hardly be recognized unless context from distant "railway" regions are incorporated into semantic parsing. Thus, incorporating context into feature maps is of great significance to achieve the desired robust high-level parsing.

Suppose a pre-trained CNN outputs a feature tensor $\mathbb{F}$. Take $\mathbb{F}$ as input, context aggregation module ($\texttt{CA}$) is to output $\mathbb{F}'$, in which each feature is well contextualized. Mathematically, $\mathbb{F}' = \texttt{CA}(\mathbb{F}, \Theta)$, where Θ denotes the learnable parameters for function $\texttt{CA}$. In general, features in $\mathbb{F}$ are interacted based on the parametric model ($\texttt{CA}$) and then contexts are encoded in the improved feature map $\mathbb{F}'$. Our view for a good contextualization mechanism is that features in $\mathbb{F}'$ should engage all elements in $\mathbb{F}$. More specifically, suppose an input feature map $\mathbb{F}$ has the spatial dimensionality of $n \times n$, the contextual view spanned by $\texttt{CA}$ module should also be approximately $n \times n$. In this paper, we propose to use basic convolutional kernels to achieve this goal. Different from atrous convolution [4], [37], there will be no "holes" during the convolution operations. Thus, it can incorporate every feature within the contextual window into corresponding features in $\mathbb{F}'$. In contrast, dilated (atrous) convolution can only incorporate a fraction of features within contextual windows. This issue is also pointed out by [34] as "gridding issue". As pre-trained CNN (with "skip" modification) outputs very coarse feature maps (16 × 16 for 512 × 512 images), we can easily construct a non-deep convolutional network to expand the expected range of contextual views for features in $\mathbb{F}'$.

1) Convolutional Context Network (CCN): We present its architecture in Fig. 3. As shown, several $\texttt{conv}$ blocks are chained to progressively expand the contextual view of feature maps. Dense skip layers are also used in CCN to aggregate multi-scale contexts.

2) Relation With ASPP in DeepLab [4]: ASPP aggregates multi-scale contexts by combing multiple $\texttt{CA}$ branches, each of which uses dilated $\texttt{conv}$ kernels with different stride rates to incorporate different scale contexts. Functionally, dense skip layers in CCN offer a similar way to aggregate multi-scale context. However, receptive field of feature maps $\mathbb{F}'$ in CCN are progressively expanded so that features in $\mathbb{F}'$ engage information from all the elements within the receptive field in $\mathbb{F}$.

In contrast, even though feature maps $\mathbb{F}'$ in some ASPP branches can expand a large contextual field (due to large stride rate), features in $\mathbb{F}'$ still incorporates information from very few elements in $\mathbb{F}$.

3) Relation With DAG-RNN [28]: DAG-RNN propagates local features to different regions of the image. Thus, features in $\mathbb{F}'$ are able to incorporate information from all elements in $\mathbb{F}$. By analyzing the unrolled structure of DAG-RNN, we discover that there are skip branches that connect every intermediate feature map with output layers. Thus, each intermediate features receive direct supervision signals, which explains why such a deep unrolled network can be effectively trained. The proposed dense skip layers in CCN offer a similar functionality to ease the difficulty of the network optimization. However, considering that the unrolled DAG-RNN is extremely deep (more than hundred layers), long-range context may vanish [25] during propagation, which limits its capability of contextual modeling. This issue is less likely to happen in the relatively shallower CCN.

4) Relation With CRF [15]: CRF is usually applied to class likelihood maps to contextualize local beliefs. In detail, the pairwise energy term is largely based on low-level image cues (e.g., color, gradient, position, etc.), thus it enforces

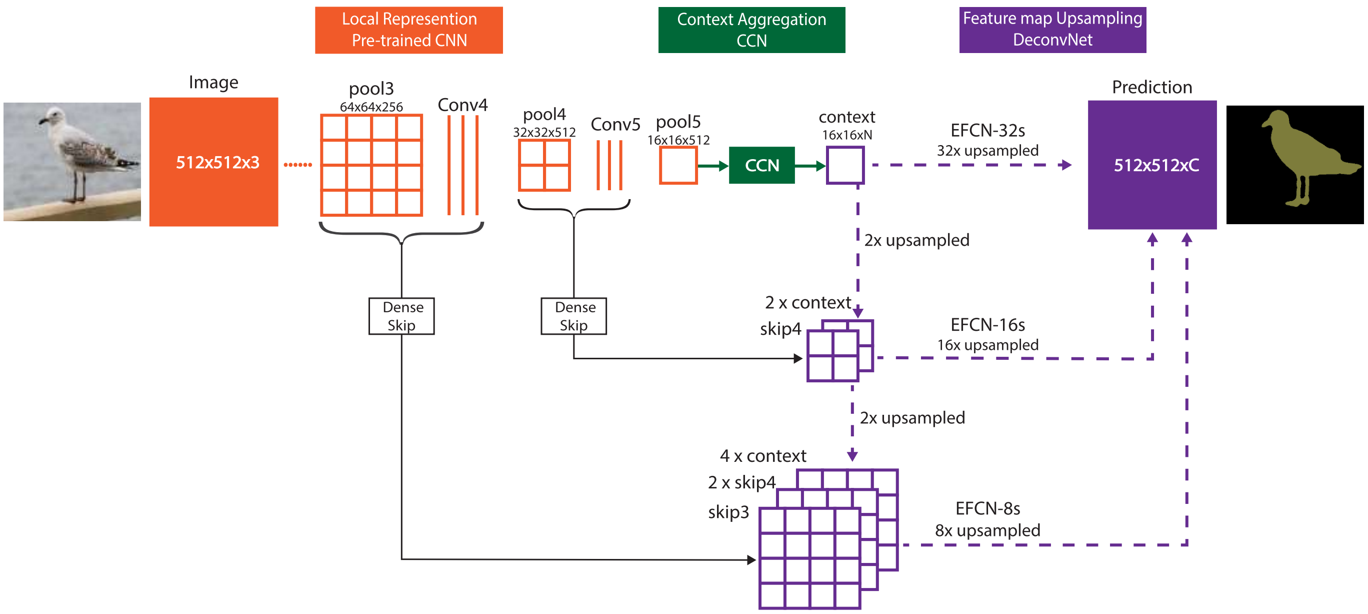


Fig. 4. Network architecture of **EFCN-xs**. Similar to the demonstration of FCN-xs, feature maps are shown as grids that reveal relative spatial coarseness. First, "dense skip" architecture is used in EFCN to retain and incorporate low-level information from pre-trained CNN, which enhances low-level visual understanding (e.g., boundary localization). Moreover, CCN is introduced to aggregate context for high-level feature maps, which brings benefits to high-level visual parsing. EFCN-4s and EFCN-2s can be trivially inferred from the above architecture demonstration. C is the cardinality of classes in the dataset and N denotes the hidden layer dimension in CCN. In order to save space, some feature maps are not displayed. Note that the detailed architecture of context network CCN can be retrieved in Fig. 3.

TABLE III

SEGMENTATION NETWORKS ARE ADAPTED TO "OUR SKIP" ARCHITECTURE. THEY DIFFERENTIATE EACH OTHER IN TERMS OF CONTEXT AGGREGATION. RESULTS ARE EVALUATED ON ADE20K DATASET

Networks	`CA` module	IOU
Baseline FCN-8s	None	29.5%
Baseline FCN-8s + CRF	CRF	31.1%
FCN-8s[27]	`fc6&fc7`	32.6%
DAG-RNN [28]	DAG-RNN	33.5%
DeepLab-v2 [4]	ASPP	34.7%
Ours	CCN	**36.4%**

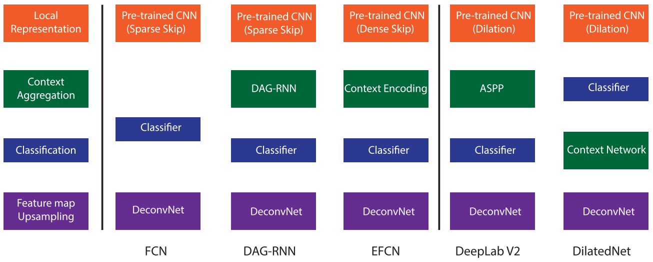


Fig. 5. Network architecture comparison. These networks differentiate each other from two aspects: (1) the way to adapt pre-trained CNN to retain spatial information; (2) the approach to aggregate context for high-level feature maps.

low-level consistency such as sharpening boundary, appearance-based smoothing, etc. Similar to ASPP [4] and DAG-RNN [28], our CCN is to contextualize high-level features so that high-level consistency can be achieved. In this sense, CRF can be placed on top of our segmentation network to further boost its parsing performance.

We carefully examined and compared these `CA` modules in a controlled experiment. For efficiency reasons, we adapt the segmentation networks based on "sparse skip" architecture. Their parsing performance on ADE20K are summarized in Table III. Here, CCN is a 8-layer convolutional network (3 × 3 kernels with 1024 channels) with dense shortcut branches. As expected, segmentation networks with `CA` modules outperforms baseline FCN, which demonstrates that context aggregation is essential to enhance feature representation. Note that even though FCN-8s [21] doesn't engage extra `CA` module, its pre-trained `fc` layers can be regarded as achieving this functionality based on its architecture difference with baseline FCN-8s. In comparison with FCN-8s, all other `CA` networks (CCN, ASPP and DAG-RNN) engage magnitude smaller parameters, but their corresponding "skip" networks achieve superior parsing performance. This result indicates that parameter sizes are not the main factor in building an effective `CA` module. Meanwhile, CCN outperforms CRF, DAG-RNN and ASPP by a significant margin, which clearly justifies our contributions as well as demonstrates its architecture superiority of feature context aggregation.

5) Comparison With Context Network in DilatedNet [37]: Architecturally, CCN has dense skip layers to aggregate multi-scale contexts, whereas context network in [37] is a plain CNN. Functionally, context network in [37] is placed on top of classification layer (c.f. Fig. 5). Thus, it models the contextual dependencies over class-likelihoods to refine local beliefs. In contrast, CCN is fed with high-level feature maps, and it aggregates and distills context to local features such that their representation capability is improved. It is an essential component in segmentation networks to enhance high-level

TABLE IV

Parsing Performance of Different Networks (c.f. Fig 5) on ADE20K Dataset

Networks	GPA	ACA	IOU	Speed (hz)
FCN(Orig) [40]	71.3%	40.3%	29.4%	10.0
FCN [27]	73.6%	44.8%	32.6%	10.0
DAG-RNN [28]	73.9%	49.0%	33.5%	9.80
DilatedNet [37]	74.1%	47.9%	33.6%	3.50
DeepLab-v2 [4]	75.1%	48.4%	35.1%	7.30
EFCN (sparse skip)	**75.8**%	**50.4**%	**36.4**%	9.80
EFCN (dense skip)	**76.2**%	**51.7**%	**37.7**%	9.00

parsing consistency. By comparing their qualitative results in Fig. 10, we clearly see that CCN noticeably outperforms context network in terms of enforcing high-level consistency.

D. EFCN

Our segmentation network is fully convolutional, and it has dense skip layers on both pre-trained CNN and CCN. Thus, we name it enhanced FCN (EFCN) based on its enhanced structure over FCN. First, "dense skip" architecture is used to retain and incorporate detailed spatial information from pre-trained CNN, which enhances the low-level visual understanding (e.g., boundary localization, small-size object detection, etc.). All skip layers in the "dense skip" network are parameterized with the architecture introduced in Section III-A. Moreover, CCN is introduced to aggregate context for high-level feature maps, which brings benefits to high-level visual parsing. Examples in Fig. 1 qualitatively illustrate such benefits to the resulting parsing maps. The network architecture of EFCN-xs is shown in Fig. 4.

It's important to mention that our contributions in this paper are orthogonal to many recent techniques that advance segmentation performance. For example, Zhao *et al.* [38] explore to leverage global scene information to improve parsing performance. Wu *et al.* [35] exploit to improve the network architecture of pre-trained CNN. Wang *et al.* [34] propose to replace bilinear up sampling with convolution to preserve detailed information.

In this paper, we follow the definition of [28] and [29] to illustrate the architecture of segmentation networks. Then, we compare the architecture of EFCN with state-of-the-art segmentation networks in Fig. 5. These networks differentiate each other from two aspects: (1) the way to adapt pre-trained CNN ("sparse skip", "dense skip" or "dilation") to retain detailed spatial information; (2) the approach to aggregate context for high-level feature maps. In Table IV, we list their parsing performance on ADE20K dataset [40]. EFCN outperforms all other networks by a significant margin, which justifies the merits of our contributions.

IV. Implementation Details

A. Network Setup

Following the functional modules of segmentation networks defined in [28] and [29] (c.f. Fig. 5), we present the detailed network architecture of our EFCN.

- **Local Representation.** We use truncated VGG-16 [31] (pre-trained on ImageNet [26]) as our local representation module. In detail, layers after `pool5` are discarded. Given an input image with size 512×512, it will output feature tensor $\mathbb{F}$ with dimensionality of $16 \times 16 \times 512$. Besides, we adapt pre-trained CNN based on the proposed "dense skip" architecture to retain detailed low-level information.
- **Context Aggregation.** Our CCN is an 8-layer 3×3 (or 6-layer 5×5) convolutional network with dense skip layers (c.f. Fig. 3), and the hidden dimension is set to 1024 (or 512). Thus, feature in $\mathbb{F}'$ is able to incorporate context from entire image.
- **Up sampling.** We use convolution transpose (deconvolution) kernels [21] to perform up sampling operation.

B. Class-Weighted Loss

To distribute more attention for infrequent classes, we modulate the pixel-wise loss according to its rareness magnitude as in [28]. This practice is economic and it is essential to significantly boost the parsing performance of rare classes. As demonstrated in Table I, a significant 2% IOU improvement is observed on ADE20K dataset when the class-weighted loss is used to train FCN. We follow the 85% -15% rule to determine the rare categories. Readers can refer to [28] for detailed description.

C. Training Details

Networks are trained with SGD with momentum (batch size 10). The learning rate is initialized to be 10^{-3}, and it is decayed by factor of 10 after 15 and 20 epochs (25 epochs in total). The momentum is fixed to 0.9. New parameters engaged in CCN and skip layers are randomly initialized (Gaussian distribution with variance 10^{-2}). Meanwhile, higher learning rate (10×) is used for newly-initialized parameters, i.e. CCN and dense skip layers. Images are resized to have maximum length of 512 pixels, and they are zero padded to 512×512 pixels to allow for batch processing. We randomly flip the images horizontally (on the fly) to augment the training images. The statistics (mean and variance) in batch normalization (`BN`) layer is updated after the network is converged. It is important to note that all the segmentation networks (in the controlled experiments) are trained with exactly the same settings.

D. Evaluation

Three performance metrics are used to evaluate our EFCN: Global Pixel Accuracy (**GPA**), Average Class Accuracy (**ACA**) and **IOU**. Readers can refer to [21] for mathematical definitions.

V. Architecture Evaluation on ADE20K

ADE20K [40] is a recently established large-scale dataset for ImageNet scene parsing challenge. The dataset contains 20210 training, 2000 validation and 3352 test images. Each pixel is annotated with one of 150 semantic categories including object classes as well as stuff classes (e.g., tree, sky and wall). In order to modulate the rareness weighted loss [28],

TABLE V

PERFORMANCE COMPARISON OF DIFFERENT EFCN ARCHITECTURES ON ADE20K DATASET. FOR EFFICIENCY REASON, THE HIDDEN DIMENSION OF CCN IS REDUCED TO 512, THUS A SLIGHT PERFORMANCE DISCREPANCY IS OBSERVED WHEN IT IS COMPARED WITH THAT IN TABLE IV

Networks	GPA	ACA	IOU
EFCN-32s	73.7%	48.4%	33.2%
EFCN-16s	75.4%	51.5%	36.3%
EFCN-8s	**76.0%**	**52.0%**	**37.0%**
EFCN-4s	75.8%	51.3%	36.0%

TABLE VI

PERFORMANCES (GPA, ACA, IOU) OF DIFFERENT NETWORK ARCHITECTURES (WITH DIFFERENT PRE-TRAINED CNNS) ON ADE20K DATASET

Networks	ResNet-50		
FCN-8s	74.4%	47.4%	33.9%
EFCN-8s	**77.0%**	**53.2%**	**38.1%**
Networks	ResNet-101		
FCN-8s	76.0%	50.1%	35.80%
EFCN-8s	77.7%	55.4%	39.74%
RefineNet[16]	-	-	40.20%
PSPNet[38]	80.6%	-	41.96%

those classes whose frequency is less than 1% are considered as rare. We report the results on validation images.

A. Which Skip Layers Should Be Included?

It's important to note that the skip layers from CCN (c.f. Fig. 3) are essential to aggregate multi-scale context as well as to ease network optimization difficulties (c.f. Section 3). The removal of them dramatically deteriorates (> 1% IOU) its parsing performance. Thus, dense skip layers from CCN should be kept in EFCN. Here, we only discuss which skip layers should be included from pre-trained CNN. To this end, we progressively add skip layers from pre-trained CNN and monitor the performance of their corresponding networks. For efficiency reasons, we consider adding a block of skip layers per evaluation.[3] In a similar notation to FCN-xs [27], we notate these networks as EFCN-xs. Their results are shown in Table V, where we can see that EFCN-8s performs the best among all "dense skip" architectures. Though EFCN-4s entails more skip layers, it achieves inferior segmentation results. We conjecture that the added block of skip layers in EFCN-4s (comparing to EFCN-8s) incorporates possibly noisy information that largely contributes to the degraded performance.

B. The Proposed EFCN Based on the Pre-Trained ResNets

We have shown that EFCN significantly outperforms FCN and other state-of-the-art VGG-16 based segmentation networks in Table IV, which verifies our contributions to the architecture design. In this section, we examine whether such architecture is universally beneficial to different pre-trained CNNs, especially to the recent developed powerful ResNets [12]. The recent success of residue learning [12], results in very deep pre-trained CNN. Also, the recent pre-trained deep networks (e.g., ResNet-101) have been expanded to very large receptive fields. Thus, it's interesting and also important to know whether the proposed architecture design (EFCN) is still able to bring performance benefits, if not as significant as that in VGG-16. To this end, we adapt the pre-trained ResNets (ResNet-50 [12] and ResNet-101 [12]) to FCN-8s and EFCN-8s correspondingly. Their parsing performance on ADE20K are listed in Table VI, where we can see that EFCN consistently and significantly outperforms its FCN counterpart. Although EFCN slightly lags behind PSPNet [38] that adopts different strategy (dilation), it brings a significant performance gain (∼ 4% IOU) from FCN that uses the same strategy of skip, which again validates the merits of the contributions in this paper. Furthermore, the proposed EFCN is faster than PSPNet and requires much less memory than PSPNet. This result is inspiring and interesting considering that FCN-8s adapted from deep ResNet is already strong. Knowing that the pre-trained ResNets are trained from low-resolution images, the introduced context network is essential and effective to adapt the feature maps so that they are optimized for the segmentation purpose of high-resolution images. Meanwhile, the fusion of rich-scale contextual predictions achieved by dense skip connections also contribute significantly to the performance boost. This encouraging result suggests that our contributions of this paper are orthogonal to the general research that improves pre-trained CNN architecture.

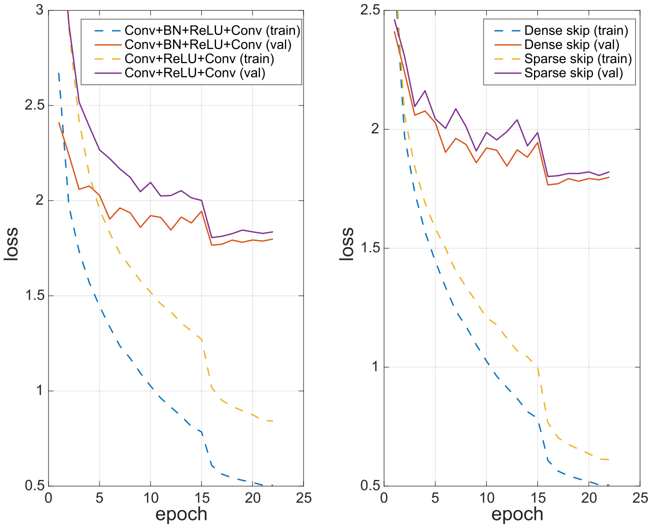


Fig. 6. Training curves of EFCN on ADE20K dataset.

C. Ablation Analysis

In this section, we firstly discuss how skip layers affect the training process of segmentation networks. In Figure 6,

[3] Feature maps are considered to be in the same block as long as they have the same spatial resolution. Take VGG-16 [31] as an example, feature maps `pool4`, `conv5_1`, `conv5_2` and `conv5_3` have the same resolution, so skip layers originating from them are deemed as in the same block.

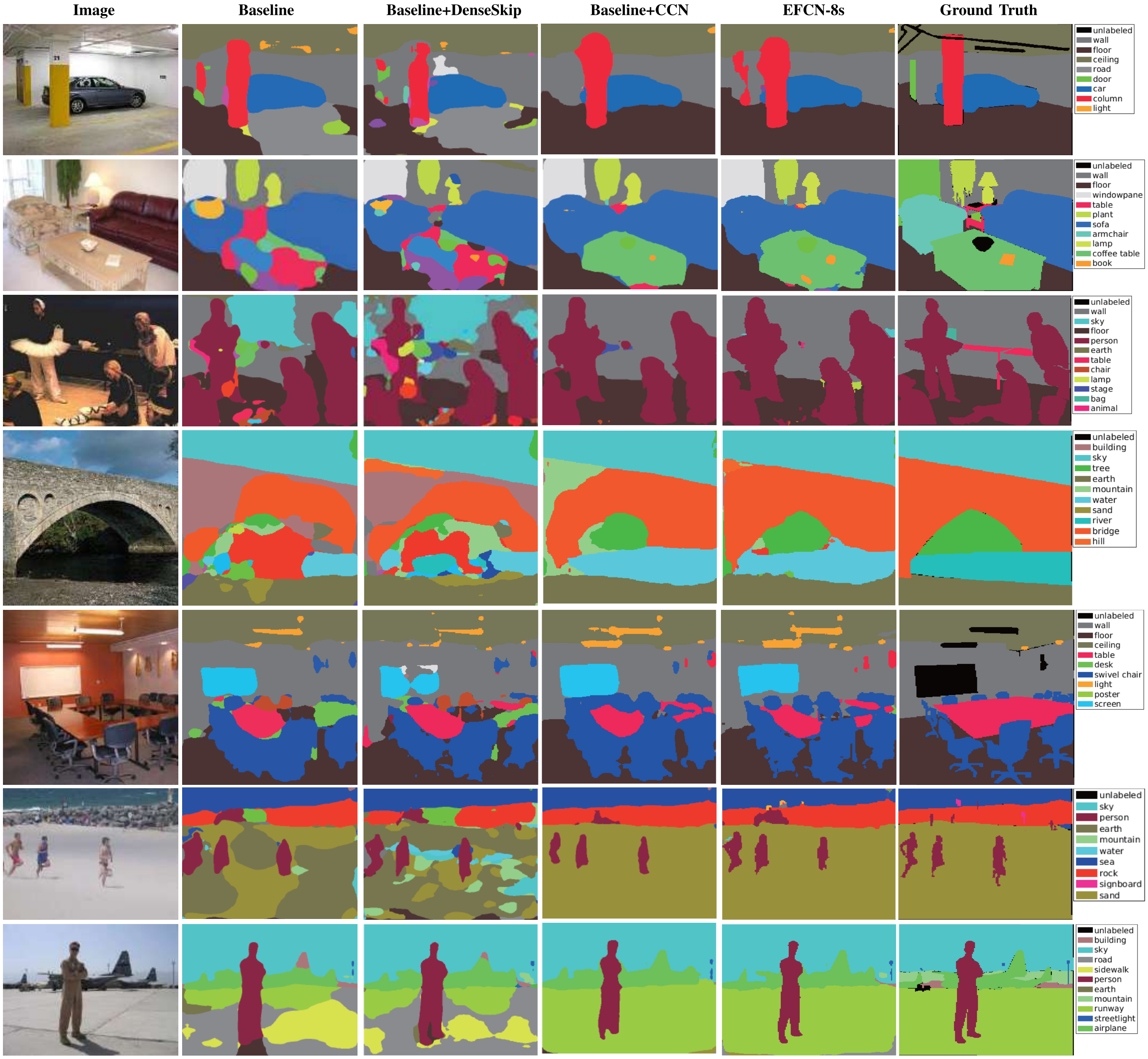


Fig. 7. Qualitative ablation analysis of the proposed segmentation network – EFCN. Images are from **ADE20K** dataset. The figure is best viewed in color with 300% zooming-in.

we compare the learning curve of training/validation loss for different segmentation network variants. As demonstrated in the left figure, batch normalization (BN) in skip layers plays a key factor in accelerating and improving training procedure. In the right figure, we clearly observe that segmentation network with "dense skip" architecture converges faster, and importantly it can reach a better local minimal. Next, we conduct step-by-step ablation studies to monitor the effectiveness of the proposed contributions. The detailed quantitative experimental results are listed in Table VII, where we can observe that each proposed contribution collectively improves the baseline segmentation network – Fully Convolutional Network (FCN) [21].

In order to interpret how those contributions improve the visual quality of label prediction maps, we demonstrate the corresponding qualitative experiment results in Figure 7. To be more specific, the proposed "dense skip" architecture helps retain detailed spatial information, which is beneficial for more accurate boundary delineation in comparison to its "sparse skip" counterpart (baseline). For example, the boundary of the 'chair' and 'person' in the 5th and 6th example has been significantly refined for "dense skip" network in contrast to its baseline ("sparse skip" architecture). In the meantime, CCN aggregates high-level contexts for feature maps, which is essential to achieve smooth and robust semantic interpretations for visually inconsistent image regions. For example, the semantic predictions for 'floor' and 'table' in the 1st and 2nd image has been significantly improved after CCN is included in the segmentation network architecture.

TABLE VII

QUANTITATIVE ABLATION ANALYSIS OF THE PROPOSED SEGMENTATION NETWORK – EFCN-8S. EXPERIMENTS ARE BASED ON THE **ADE20K** DATASET

Skip with BN	Dense Skip	CCN	IOU
✗	✗	✗	31.4%
✓	✗	✗	32.6%
✓	✓	✗	33.6%
✓	✗	✓	36.4%
✓	✓	✓	37.0%

TABLE VIII

PASCAL CONTEXT VALIDATION ACCURACIES. FOR FAIR COMPARISON, WE ONLY INCLUDE METHODS THAT USE VGG-16 AS BASE NETWORK

Methods	GPA	ACA	IOU
FCN-8s [27]	67.0%	50.7%	37.8%
DeepLab-v1 [4]	n/a	n/a	37.6%
CRF-RNN [39]	n/a	n/a	39.3%
ParseNet [19]	n/a	n/a	40.4%
UoA-Context + CRF [17]	71.5%	53.9%	43.3%
DAG-RNN + CRF [28]	73.6%	55.8%	43.7%
EFCN-8s	**74.5%**	**57.7%**	**45.0%**

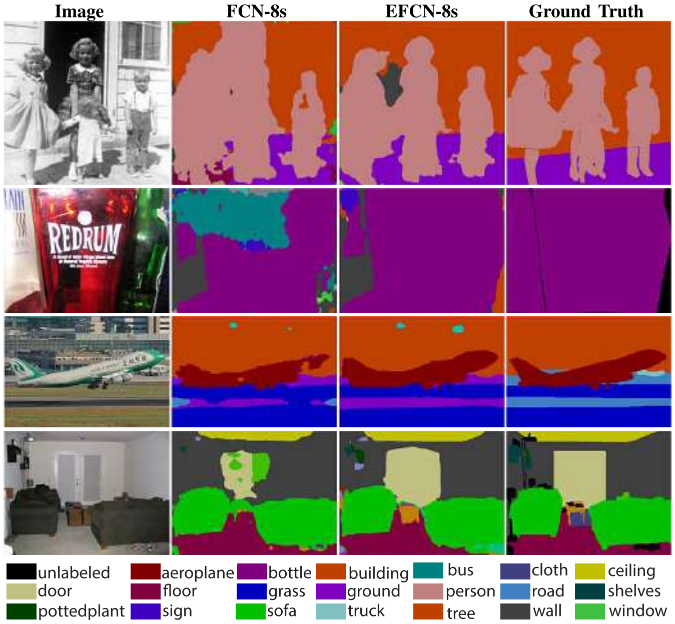


Fig. 8. Qualitative segmentation result comparison on **Pascal Context**. In each row, we show input images, unary prediction maps from FCN-8s, EFCN-8s (VGG-16) and ground truth.

VI. RESULTS ON SEGMENTATION BENCHMARKS

We evaluate the proposed EFCN on standard scene segmentation datasets. The same experimental setups (as in controlled experiments, c.f. Section IV) are used to train our segmentation network over different datasets.

ADE20K results are listed in Table IV,VI.

Pascal Context results are shown in Table VIII and qualitative segmentation result comparisons are presented in Fig. 8. Pascal Context [23] has 10103 images, out of which 4998 images are used for training. The images are from Pascal VOC 2010 datasets, and they are re-labeled as pixel-wise segmentation maps which include 540 semantic classes (including the original 20 classes). Similar to Mottaghi *et al.* [23], we only consider the most frequent 59 classes in the dataset for evaluation. Based on the rareness identification rule in [28], those classes whose frequencies are lower than 1% are identified as rare.

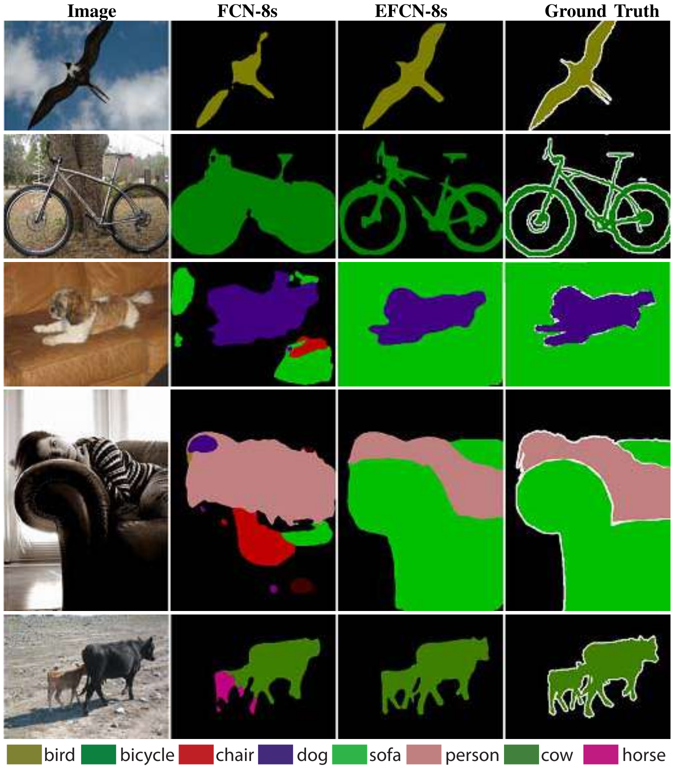


Fig. 9. Qualitative segmentation result comparison on **Pascal VOC 2012**. EFCN outperforms FCN in terms of both the low-level (first two examples) and high-level (last three examples) semantic parsing, which clearly aligns with our motivations.

TABLE IX

SUN-RGBD (37 CLASSES) TESTING ACCURACIES. WE ONLY USE RGB MODALITY IN OUR EXPERIMENTS. ALL METHODS USE VGG-16 AS BASE NETWORK, AND ALL OTHER REPORTED RESULTS ARE COPIED FROM [1]

Methods	GPA	ACA	IOU
FCN [21]	68.18%	38.41%	27.39%
DeconvNet [24]	66.13%	33.28%	22.57%
SegNet [1]	72.63%	44.76%	31.84%
DeepLab [4]	71.90%	42.21%	32.08%
EFCN-8s	**76.90%**	**53.46%**	**40.74%**

SUN-RGBD Results are reported in Table IX. SUN-RGBD [32] contains images from NYU depth V2 [30], Berkeley B3DO [14], SUN3D [36] as well as the newly captured images. It has 10335 images in total, out of which 5285 images are used for training. The rareness frequency threshold is fixed to 2.5% based on the 85%-15% rule [28].

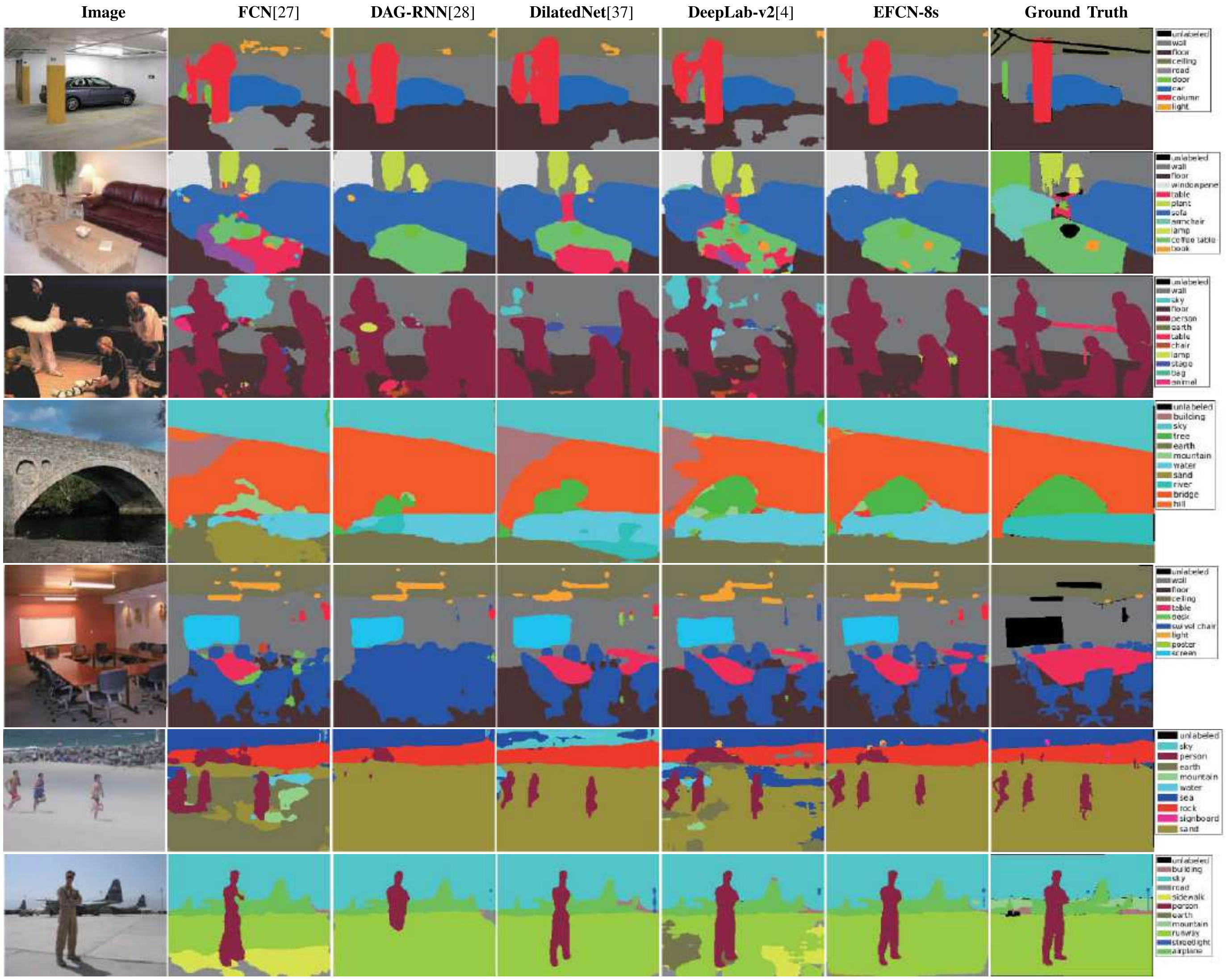


Fig. 10. (Best viewed in color with 300% zooming-in) Qualitative comparison of parsing maps on ADEK20K dataset. DAG-RNN [28] is good at enforcing high-level consistency (first three examples). Due to the adopted "dilation" architecture, DeepLab-v2 [4] and DilatedNet [37] perform well at localizing boundaries (last three examples). EFCN performs well in both the low-level and the high-level parsing.

We follow previous literatures [1] to consider 37 classes for evaluation. Note that we only use RGB modality as input.

Pascal VOC 2012 results are presented in Table X and qualitative segmentation result comparisons with FCN [27] are shown in Fig. 9. Pascal VOC 2012 originally contains 2913 train and validation images, 1456 testing images. Images have approximately 375×500 pixels, each of which belongs to one of the pre-defined object or background categories. We follow [21] to augment the training set from [10]. Thus, we end up with having 12031 training images. The rareness frequency threshold of is set to 1.5%. We have tested four models: (1) EFCN-8s without any pre-training or post-processing. (2) Post-processing EFCN-8s by CRF [15], which brings 0.7% improvement on IOU. (3) EFCN-8s pre-trained on Microsoft COCO dataset [18], improving 2.4% on IOU. (4) EFCN-8s pre-trained on Microsoft COCO dataset and then post-processing by CRF. We submit our parsing maps of testing images to the evaluation servers to report our results.

A. Result Summary

EFCN demonstrates significantly better quantitative parsing performance than FCN on all segmentation benchmarks. As the evaluation datasets cover wide range of scenarios that include object segmentation [6], outdoor scene parsing [23], [40] as well as indoor scene labeling [40], we can conclude that EFCN is a versatilely better segmentation network architecture than FCN. By comparing their qualitative parsing results in Fig. 1, Fig. 8, Fig. 9 and Fig. 10, we observe that EFCN outperforms FCN in terms of both the high-level and low-level semantic parsing, which aligns with our motivations. The significantly enhanced quantitative performance and the noticeable improved visual quality of parsing maps justifies the contributions of our proposed EFCN in this paper.

EFCN vs State-of-the-Arts: EFCN also outperforms other recently developed parsing networks including DilatedNet [37], DAG-RNN [28], DeepLab [4], etc. By comparing their parsing maps in Fig. 10, it is not difficult to find that

TABLE X

PASCAL VOC 2012 TESTING ACCURACIES. FOR FAIR COMPARISON, WE ONLY INCLUDE METHODS THAT USE VGG-16 AS BASE NETWORK

Methods	IOU
FCN-8s [27]	67.5%
Zoom-out [22]	69.6%
PixelNet [2]	69.7%
ParseNet [19]	69.8%
DeepLab-v1 + CRF [3]	71.6%
CRF-RNN [39]	72.0%
DeconvNet [24]	72.5%
DilatedNet + COCO pre-trained[37]	73.5%
DPN [20]	74.1%
UoA-Context + CRF[17]	75.3%
LRR [8]	74.7%
EFCN-8s	75.2%
EFCN-8s + CRF	75.9%
EFCN-8s + COCO pre-trained	77.6%
EFCN-8s + COCO pre-trained +CRF	**78.3**%

EFCN achieves the best parsing performance both in low-level and in high-level parsing.

In the setting of VGG-16, the unary predictions of EFCN-8s outperforms state-of-the-arts by a large margin on ADE20K, Pascal Context and SUN RGB-D datasets. On Pascal VOC 2012, EFCN-8s also achieves very competitive segmentation performance among all top methods.

VII. CONCLUSION

In this paper, we explore ways to develop effective as well as efficient segmentation networks that can perform well on both the low-level and the high-level semantic parsing. To this end, we first discuss and compare two popular adaptation approaches to retain detailed spatial information from pre-trained CNN -"dilation" and "skip". By modifying the parameterization of skip layers, "our skip" network outperforms the conventional "skip" network by a noticeable margin. Furthermore, we propose "dense skip" architecture to efficiently retain rich set of low-level information. For high-level semantic parsing, we propose a convolutional context network (CCN) to aggregate context for high-level feature maps so that their representation capability can be improved. The resulting network architecture (EFCN) performs competitively well on standard scene segmentation datasets. Without bells and whistles, the proposed EFCN achieves state-of-the-arts performance on ADE20K, Pascal Context, SUN RGB-D and VOC 2012 datasets.

ACKNOWLEDGMENT

This research was supported by Singapore Ministry of Education Academic Research Fund Tier 1 RG 123/15 Grant no: 2015-T1-002-140. This research was carried out at the Rapid-Rich Object Search (ROSE) Lab, School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. (2015). "SegNet: A deep convolutional encoder-decoder architecture for image segmentation." [Online]. Available: https://arxiv.org/abs/1511.00561

[2] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan. (2016). "PixelNet: Towards a general pixel-level architecture." [Online]. Available: https://arxiv.org/abs/1609.06694

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. ICLR*, 2015, pp. 1–14.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2016). "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915

[5] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2393–2402.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[8] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 519–534.

[9] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[10] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 297–312.

[11] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 447–456.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 1–11.

[14] A. Janoch *et al.*, "A category-level 3D object dataset: Putting the kinect to work," in *Consumer Depth Cameras for Computer Vision*. London, U.K.: Springer, 2013, pp. 141–165.

[15] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. NIPS*, 2011, pp. 109–117.

[16] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, p. 5.

[17] G. Lin, C. Shen, A. van dan Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.

[18] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[19] W. Liu, A. Rabinovich, and A. C. Berg. (2015). "ParseNet: Looking wider to see better." [Online]. Available: https://arxiv.org/abs/1506.04579

[20] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1377–1385.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[22] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3376–3385.

[23] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 891–898.

[24] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[25] R. Pascanu, T. Mikolov, and Y. Bengio. (2012). "On the difficulty of training recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1211.5063

[26] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[27] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[28] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "DAG-recurrent neural networks for scene labeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2016, pp. 3620–3629.

[29] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "Scene segmentation with DAG-recurrent neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1480–1493, Jun. 2018.

[30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 746–760.

[31] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[32] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 567–576.

[33] F. Visin *et al.*, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2016, pp. 41–48.

[34] P. Wang *et al.* (2017). "Understanding convolution for semantic segmentation." [Online]. Available: https://arxiv.org/abs/1702.08502

[35] Z. Wu, C. Shen, and A. van den Hengel. (2016). "Wider or deeper: Revisiting the resnet model for visual recognition." [Online]. Available: https://arxiv.org/abs/1611.10080

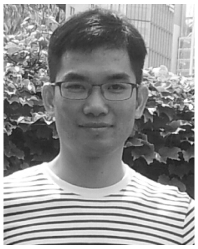
[36] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1625–1632.

[37] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[39] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1529–1537.

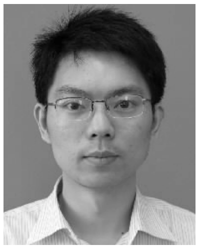
[40] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. (2016). "Semantic understanding of scenes through the ade20k dataset." [Online]. Available: https://arxiv.org/abs/1608.05442



Bing Shuai received the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2018. He is currently a Research Scientist at Amazon AWS Rekognition. His research interests include computer vision (specifically in) scene segmentation, video segmentation, and video action detection.

Henghui Ding received the B.E. degree from Xi'an Jiaotong University, China, in 2016. He is currently pursuing the Ph.D. degree with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and machine learning.

Ting Liu received the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Algorithm Engineer at Alibaba AI Labs. His research interests reside in visual tracking and object detection.



Gang Wang received the B.Eng. degree in electrical engineering from the Harbin Institute of Technology and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. He was an Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. He had a joint appointment at the Advanced Digital Science Center, Singapore, as a Research Scientist, from 2010 to 2014. He is currently a Researcher/Senior Director and a Distinguished Scientist at Alibaba AI Labs. He was a recipient of MIT Technology Review Innovators Under 35 Award (Asia). He is an Area Chair of ICCV 2017 and CVPR 2018. He is an Associate Editor of TPAMI.

Xudong Jiang (M'02–SM'06) received the B.Eng. and M.Eng. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983 and 1986, respectively, and the Ph.D. degree from Helmut Schmidt University, Hamburg, Germany, in 1997, all in electrical engineering. From 1986 to 1993, he was a Lecturer with UESTC, where he received two Science and Technology Awards from the Ministry for Electronic Industry of China. From 1998 to 2004, he was with the Institute for Infocomm Research, A*STAR, Singapore, as a Lead Scientist, and the Head of the Biometrics Laboratory, where he developed a system that achieved the most efficiency and the second most accuracy at the International Fingerprint Verification Competition in 2000. He joined Nanyang Technological University (NTU), Singapore, as a Faculty Member, in 2004, where he has also served as the Director of the Centre for Information Security from 2005 to 2011. He is currently a tenured Associate Professor with the School of Electrical and Electronic Engineering, NTU. He holds seven patents and has authored over 150 papers with over 30 papers in the IEEE journals, including nine papers in the IEEE TIP, five papers in the IEEE TPAMI, and three papers in the IEEE TSP. His research interests include signal/image processing, pattern recognition, computer vision, machine learning, and biometrics. He has been an Elected Voting Member of the IFS Technical Committee of the IEEE Signal Processing Society, and he has served as Associate Editors for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE SIGNAL PROCESSING LETTERS, and *IET Biometrics*.