Filling the Gaps in Atrous Convolution: Semantic Segmentation With a Better Context

LIYUAN LIU¹, YANWEI PANG¹, SYED WAQAS ZAMIR², SALMAN KHAN², FAHAD SHAHBAZ KHAN², AND LING SHAO²

¹Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
²Inception Institute of Artificial Intelligence, Abu Dhabi 999041, United Arab Emirates
Corresponding author: Yanwei Pang (pyw@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61632018, and in part by the Science and Technology Innovation 2030: the Key Project of Next Generation of Artificial Intelligence under Grant 2018AAA01028.

ABSTRACT The main challenge for scene parsing arises when complex scenes with highly diverse objects are encountered. The objects not only differ in scale and appearance but also in semantics. Previous works focus on encoding the multi-scale contextual information (via pooling or atrous convolutions) generally on top of compact high-level features (i.e., at a single stage). In this work, we argue that a rich set of cues exist at multiple stages of the network, encapsulating low, mid and high-level scene details. Therefore, an optimal scene parsing model must aggregate multi-scale context at all three levels of the feature hierarchy; a capability that lacks in state-of-the-art scene parsing models. To address this limitation, we introduce a novel architecture with three new blocks that systematically aggregate low, mid and high tier features. The heart of our approach is a high-level feature aggregation module that augments sparsely connected atrous convolution with dense local and layer-wise connections to avoid gridding artifacts. Besides, we employ a novel feature pyramid augmentation and semantic refinement unit to generate low- and mid-level features that are mixed with high-level features at the decoder. We extensively evaluate our proposed approach on the large-scale Cityscapes and ADE2K benchmarks. Our approach surpasses many latest models on both datasets, achieving mean intersection-over-union (mIoU) scores of 80.5% and 44.0% on Cityscapes and ADE20K, respectively.

INDEX TERMS Image processing, neural networks, semantic segmentation, supervised learning.

I. INTRODUCTION

Given an image, the goal of semantic segmentation is to assign a category label to each pixel [1], [2]. It is a challenging problem with numerous real-world applications, including autonomous driving, satellite imaging and scene understanding [3]. The semantic segmentation task simultaneously deals with two sub-problems: classification and localization. However, both these sub-tasks are contradictory in nature and largely influence the design principles of the segmentation model. In the case of classification, the models aim to learn higher-level representations to capture the global context and therefore need to be invariant to local image details such as in-plane transformations and deformations. [4] Whereas such abstraction of spatial information is undesirable for the localization task, for which the models need to resolve fine-grained local information with pixel-level accuracy. [5] These two conflicting aspects of the problem mean that a range of complementary features are desired to build a robust model that balances the two extremes based on the given context.

The recent well-known models such as PSPNet [7] and DeepLab [8], [9] only perform multi-level contextual learning at a single stage within the deep network hierarchy. Specifically, PSPNet [7] applies pooling operations with different sub-sampling rates, all arranged in parallel, to capture context information. Their pyramid pooling module works only on last convolutional layer features, that generally lacks local scene details. The Atrous Spatial Pyramid Pooling (ASPP) module in DeepLabv2 [8] & v3 [9] applies parallel atrous convolutions with different dilation rates to extract multi-scale context information, however, they also work on a high-level feature representation. Similarly, [10] uses dense
connections between convolution layers with multiple dilation rates and only operates on later-stage features in the deep network. Besides, all the above-mentioned approaches use a direct feature upsampling for final prediction. In this paper, we argue that a simple upsampling operation is insufficient to fully recover the fine details and is sub-optimal to combine multi-stage (mainly low and mid-level) features for final prediction (see Figure 1).

One alternative to resolve the above-mentioned limitation was proposed in ‘U-Net’ architecture [6], [11] that gradually fuses high-level features from top layers with low-level features from initial layers. Despite being intuitive, this design does not consider both mid-level features and multi-scale context on low and high-level representations. Similar architecture designs have been proposed in GCN [12], DeepLabv3+ [13], PANet [14] and ExcuseNet [15] but with a similar limitation. Some recent models, such as RefineNet [16] and DFN [17], not only design a top-down smooth network but also apply a bottom-up border network to refine the final output. However, the low-level features in these networks are actually converted to mid-level features before they are used for refinement. This results in loss of sharp details (e.g., around boundaries) and mistakes for small objects. In order to address these issues, we propose to jointly integrate low-level features in the network, along-side mid and high-level cues.

A. CONTRIBUTIONS

We propose and approach with the following main contributions: (a) We propose a feature pyramid based augmentation module to efficiently generate refined low-level features to preserve the local details. (b) For mid-level multi-scale feature fusion, we propose a semantic refinement unit that combines a diverse set of features from the network encoder. (c) The central component of our model is a high-level context aggregation block. It is based on the insight that dilated convolution expands kernel size by interleaving its weights with zeros, that equates to dropping the intermediate activations in the input feature map. To alleviate this problem, we propose to combine the strengths of dilated (sparse) and wider (dense) kernels, that eventually enhance the discriminative power of the network and avoids unfairly neglecting the local information, as is the case with atrous convolution. We perform experiments on Cityscapes and ADE20K datasets. Our approach achieves superior results compared to existing methods on both datasets. Specifically, it achieves 80.5% and 44.0% mIoU scores on Cityscapes and ADE20K datasets, respectively, outperforming the best reported results in [7], [17].

II. RELATED WORK

The seminal work of Long et al. [18] on fully convolutional networks (FCNs) exceeded the performance of all previous hand-crafted [19], [20] and learning-based segmentation methods [21]–[24]. FCN based segmentation frameworks typically employ (pretrained) image classification networks as backbone that use deep pooling hierarchy to increase the receptive field for better encoding of image semantics. However, the repeated use of pooling layers and convolutions with striding operations drastically reduce the spatial dimensions of the feature maps. When such features maps are upsampled back to full resolution at the output layer, the produced segmentation results appear coarse with deteriorated object boundaries [18].

To alleviate this problem, Chen et al. [25] propose to remove the downsampling operator from the last few pooling layers of the backbone and instead use convolution with upsampled filters (aka. atrous convolution [26]), thus producing dense features maps. The atrous convolution produces feature maps in which all units have the same receptive field size and therefore incorporate the objects’ context at a single scale, which is a limitation as objects exist at multiple scales in real-world images.

Inspired from spatial pyramid pooling (SPP) [27], PSPNet fuses features at different pooling pyramid scales. The ASPP network [8] applies atrous convolution with multiple dilation rates in parallel, thus effectively increasing the receptive field as well as capturing object and image context at multiple scales. In certain applications, e.g., autonomous driving and land cover mapping the receptive field of neurons is required to be really large, which ASPP [8] can only achieve by using large dilation rates. However, with the increased dilation rate the atrous convolution starts to lose its modeling power and gradually becomes less effective.

DenseASPP [10] has been proposed to incorporate wider context in the ASPP block, especially in cases where dilation rates are high. Although DenseASPP improves context, one limitation of such design is that only the high-level features from the later layers in a deep CNN are used to model contextual information. This limitation makes [10] and previous network designs attentive to local scene details. One promising way to counter this effect is through a encoder-decoder
The overall framework. (a) An overview of the entire network architecture. (b) LKD block contains a large kernel and a dilation convolution branch, connected in parallel. (c) Details of proposed feature pyramid module. (d) The refinement unit. Our proposed design combines low (c), mid (d) and high-level (b) features to obtain a refined semantic labeling map.

structure [11], [29]. The encoder-decoder [11], [12], [15], [16], [29]–[32] networks acquire the semantic detail in the encoder branch and try to recover sharp object boundaries in the decoder branch.

### A. OUR APPROACH

The existing network architectures lack a multi-stage and systematic combination of high, mid and low level image features. Our proposed model design demonstrates that none of the three levels is independently sufficient for optimal multi-scale feature fusion. Specifically, our model introduces novel network designs for all the three context-aware blocks in the proposed model that leads to significantly better performance. The proposed approach outperforms existing methods, mostly employing high-level features, on both Cityscapes and ADE2K datasets.

### III. APPROACH

Figure 2 illustrates the core components of our framework. These include: (1) large kernel paralleled with dilation (LKD) block, (2) refinement unit, and (3) low-level feature pyramid (LFP). Next, we first describe the functionality of each of these individual components. Later we show how they fit together in overall architecture to complement each other with different type of information, thus providing us with accurate semantic segmentation results.

### A. LKD BLOCK

The dilated convolution (‘algorithmé á-trous’) provides an excellent mechanism to expand the receptive field without any loss of spatial resolution [26]. It enlarges the convolution kernel with a dilation factor $d$ such that the kernel weights are separated by $d - 1$ zeros, hence á-trous (with holes). However, dilated convolution suffers from gridding artifacts that hamper the performance for segmentation tasks due to two main reasons. First, the receptive field for each kernel has a checkerboard shape which sub-samples an input unit without considering its neighborhood information. This results in loss of local information especially when dilation rates are high. Second, the repeated zeroed patterns in the dilated kernel cause the neighboring points to be computed from mutually exclusive sets of input units. Figure 1 has shown this problem and highlights the importance of local information.

Some recent approaches aim to achieve ‘degridding’ in the dilated convolution outputs. Stacking several dilated convolution layers with fixed [33] or varied [35] dilation rates have shown to be useful but at the cost of additional computation complexity. An alternative with low computation cost was proposed in [34] that simply smooths out the inputs to make them inter-dependent. In contrast to these approaches, our motivation is different: since the local context can be helpful or misleading for the dense local predictions, we want to systematically decide whether and how to use it optimally beside more wider contextual cues.
We, therefore, propose to concatenate the feature responses from both the sparse and dense kernels\(^1\) with a large receptive field. It not only enhances the effective receptive field, but most importantly preserves the complementary neighborhood aware and neighborhood agnostic features that prove very useful in resolving local ambiguities depending on the given case (see Figure 1 as an example). Consider an input feature \(x\) and a convolutional kernel \(w\), the output feature map can be given by:

\[
y(i) = (x * w)(i) = \sum_k x(i + k)w(k). \tag{1}
\]

Beside \(y(i)\), we also obtain a complimentary feature map using a sparse kernel that is dilated with a factor \(d\):

\[
\hat{y}(i) = (x *_d w)(i) = \sum_k x(i + d \cdot k)w(k). \tag{2}
\]

The output of the LKD block is the concatenation of both feature responses i.e., \(o = y \oplus \hat{y}\). Note that dense convolutions with large kernel convolution are typically expensive, therefore we use the asymmetric convolution via spatial factorization scheme proposed in [28] to reduce the computation cost of large kernel operation.

**B. REFINEMENT UNIT (RU)**

In a task like semantic segmentation, features from all network layers are useful in providing efficient dense predictions. Therefore, several previous works [13], [15]–[17] combine low-level features from earlier layers that contain abundant spatial information with high-level features from deeper layers that are semantically enriched. On one hand, high-level semantic features are helpful in recognizing image regions’ labels. On the other hand, low-level features are good at generating sharper object boundaries for high-resolution predictions.

Traditional encoder-decoder architectures (like U-Net [11]) use skip-connections to directly pass features from the lower layers of the encoder to the corresponding higher layers of the decoder. In this paper, we introduce refinement units in the skip pathways to filter the flow of information. The refinement unit effectively exploits information from encoder layers for each spatial region in a top-down manner. Simply put, at each layer of the encoder, the refinement unit allows every subsequent high-level layer to provide contextual feedback in refining the relatively low-level features of the current layer, which are then passed on to the decoder as the mid-level features. Figure 2(d) illustrates the architecture of the refinement unit.

The refinement unit takes as input the encoder feature maps \(f^i_e\) and \(f^{i+1}_e\), where \(j = \{1, \ldots, N\}\) and \(N\) being the number of remaining encoder blocks that lie onward from the \(i^{th}\) block. The encoder features \(f^i_e\) have high spatial resolution but less contextual information. Whereas \(f^{i+1}_e\) features maps are of progressively low-resolution but with abundant semantic information. For the sake of simplicity, let us assume \(j = 1\).

The task of refinement unit is to produce rich contextual features by combining both \(f^i_e\) and \(f^{i+1}_e\) maps. The two input feature maps \(f^i_e\) and \(f^{i+1}_e\) may differ in spatial resolution and numbers of channels. Therefore, we first match their channel dimensions by applying a matching function \(F(\cdot)\). The matching function is a small sub-net comprising of \(3 \times 3\) convolution followed by batch normalization and a ReLU non-linear activation. Next we upsample \(f^{i+1}_e\) by a factor of two to match its spatial dimensions with \(f^i_e\) map. Finally, we perform an element-wise multiplication on these two feature maps in order to produce the refined features \(R^{i+1}_f = F(f^i_e) \otimes F(f^{i+1}_e)\).

In the case when high-level features come from multiple encoder blocks (i.e., \(j = 1, \ldots, N\)), we add the individual refined feature maps together to obtain the final feature responses, as follows:

\[
R_f = \sum_{j>i} F(f^i_e) \otimes F(f^{i+1}_e). \tag{3}
\]

As shown in Figure 2(d), the refined features maps \(R_f\) are processed with some convolutional layers (followed by batch normalization and ReLU non-linearity) to further enrich them with mid-level semantics. Finally, these feature maps are passed to the decoder.

**C. LOW-LEVEL FEATURE PYRAMID (LFP)**

Natural images exhibit large scale variation of objects. A segmentation method that works well for a particular object scale might perform poorly at other scales. For instance, making predictions on images with large resolution is beneficial for small objects and vice versa. To address the problem of scale variation, several solutions have been proposed such as combining features from shallower layers with deeper layers [13]–[17] to particularly deal with small objects, dilated convolutions [8], [10], [26] to increase receptive field, independently performing dense prediction at different resolutions to treat objects at each scale differently.

Besides our key contributions described in previous sections on aggregating context of objects at multiple scales, in this section we introduce a novel low-level feature pyramid (LFP) that complements the features in the decoder to further improve the performance of the network. The LFP module takes the original image as input and reduces its spatial dimension by the factor of 4 using bilinear interpolation. Different from the previous works [36], [40], our LFP does not contain many convolution operations, which greatly guarantees to incorporate the true low-level information of the image. The reduced-dimension image is processed with a light-weight convolutional network to obtain a feature pyramid. From this tensor we form a pyramid with three different scales: (a) original-sized low-level features, (b) low-level features downsampled by the factor of 2 using a \(3 \times 3\) convolution with \(stride = 2\), and (c) low-level features downsampled by

---

\(\text{1}\) sparse kernels have padded zeros in between weights, dense kernels have all tunable weights.
With dilated convolution we especially in the case of high dilation rate (e.g. 6), accumulate the influence on each neuron from some of the far away neurons. On the other hand, we aggregate responses for each neuron from all neurons in its local neighborhood with large kernel convolution. Together the large kernel branch and the dilated convolution branch aggregate local and non-local context. We connect each LKD block with all subsequent LKD blocks in a feed-forward fashion by using dense connections, which is different from ASPP module [8] that has four independent parallel branches. Therefore, in our network the features from different LKD blocks are well integrated. Moreover, we use the global average pooling operation to obtain a descriptor that incorporates the global context. We upsample this descriptor in order to concatenate it with the output of the last LKD block.

Finally we have a decoder in the network that receives high-level features from the last LKD block, low-to-mid level features from the refinement unit and low-level features from the low-level feature pyramid module. The decoder effectively combines all these features to make the final dense prediction of semantic segmentation. Figure 3 illustrates outputs at different stages in the decoder. We can see that the extracted features become progressively more precise and objects that are mixed together are gradually separated, resulting in segments with well defined contours.

IV. EXPERIMENTS

The proposed method aggregates low, mid and high level features at multiple scales in order to provide precise segmentation results on images having small, medium and large size objects depicting scale variations. To show the effectiveness of our framework, we carry out comprehensive experiments on two large-scale diverse datasets: Cityscapes [43] and ADE20K [44]. The quantitative performance is measured in terms of mean class-wise intersection-over-union (mIoU). Next we briefly describe each of these benchmark datasets.

A. DATASETS

1) CITYSCAPES [43]

This dataset is particularly developed for semantic understanding of urban street scenes. It is a large-scale diverse dataset that is captured in 50 different cities and comprised of 5,000 high resolution (2048 × 1024) pixel-level annotated street images of which 2,975 are for training, 500 for validation and 1,525 for testing. Note that in our experiments, we consider only those 19 categories that have finely annotated labels.

2) ADE20K [44]

It contains 20,210 training, 2,000 validation and 3,000 test images with 150 stuff/object category labels. It is a challenging dataset in which the frequency of objects in images ranges from 5 to 273.

B. IMPLEMENTATION DETAILS

The backbone of our network is ResNet-101 pretrained on ImageNet-1k. Following the works of [9], [28], we replace
TABLE 1. Ablation study on dilation rates and large kernel sizes.

| dilation conv rate set | large kernel k set | mIoU (%) |
|------------------------|--------------------|----------|
| (6.9,12.18)           | (7,11,13,19)       | 77.04    |
| (6.9,12.18)           | (5,9,11,17)        | 76.90    |
| (6.12,18.24)          | (5,11,17,23)       | 77.21    |

Ablation study w.r.t. the effect of dilation rates and large kernel sizes on Cityscapes val set. Note that all results are computed without decoder.

TABLE 2. Ablation study on different layout of large kernel and dilation convolution branches.

| dilation conv rate set | connection type | large kernel k set | mIoU (%) |
|------------------------|-----------------|--------------------|----------|
|                        | cascaded        |                    | 76.46    |
|                        | parallel        |                    | 76.95    |
|                        | dense           |                    | 76.77    |
|                        | dense           |                    | 76.14    |
|                        | dense           |                    | 77.38    |

Ablation study w.r.t. the effect of different layout of large kernel and dilation convolution branches on Cityscapes val set. We use {6.9,12.18} rates for dilation convolutions and {5.9,11.17} kernel sizes for large kernels. Note that all results are computed without decoder.

TABLE 3. Ablation study of different modules of the network.

| SCM | RU | LFP | mIoU(%) |
|-----|----|-----|---------|
| Baseline | ✓  |     | 76.77   |
| Ours  | ✓  | ✓   | 78.02   |
|       | ✓  | ✓   | 79.31   |

Ablation study: effect of semantic consolidation module (SCM), refinement unit (RU) and low-level feature pyramid (LFP). For fair comparison, all results are obtained using the same backbone (ResNet-101). Performance is evaluated on Cityscapes val set with single-scale input. Our final approach provides a significant improvement of 2.54% in terms of mIoU, compared to the baseline.
TABLE 4. Per-class comparison on Cityscapes test set.

| Methods  | road  | walk | build | wall  | fence | pole  | light | sign  | veg   | terrain | sky   | person | rider | car   | truck | bus   | train | mbike | bake | mIoU(%) |
|----------|-------|------|-------|-------|-------|-------|-------|-------|-------|--------|-------|--------|-------|-------|-------|-------|-------|-------|-------|-------|
| FCN      | 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   |
| DilatedNet | 97.6  | 79.2 | 89.9  | 37.3  | 47.6  | 53.2  | 58.6  | 65.2  | 91.8  | 69.4   | 93.7  | 78.9   | 55.0  | 93.3  | 45.5  | 53.4  | 47.7  | 52.2  | 66.0  | 67.1   |
| DeepLabv2 [8] | 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 [16] | 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   |
| GCN [12] | -     | -    | -     | -     | -     | -     | -     | -     | -     | -      | -     | -      | -     | -     | -     | -     | -     | -     | -     | 76.9   |
| DUC [35]    | 98.5  | 85.5 | 92.8  | 58.6  | 55.5  | 65.0  | 73.5  | 77.9  | 93.3  | 72.0   | 95.2  | 84.8   | 68.5  | 95.4  | 70.9  | 78.8  | 68.7  | 65.9  | 73.8  | 77.6   |
| PSpNet [7]  | 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  | 93.8  | 68.2  | 79.5  | 73.8  | 69.5  | 77.2  | 78.4   |
| Pan [14]    | -     | -    | -     | -     | -     | -     | -     | -     | -     | -      | -     | -      | -     | -     | -     | -     | -     | -     | -     | 78.6   |
| AAP [45]    | 98.5  | 85.6 | 93.0  | 53.8  | 59.0  | 65.9  | 75.0  | 78.4  | 93.7  | 72.4   | 95.6  | 84.6   | 70.5  | 95.9  | 73.9  | 82.7  | 76.9  | 68.7  | 76.4  | 79.1   |
| DFN [17]    | -     | -    | -     | -     | -     | -     | -     | -     | -     | -      | -     | -      | -     | -     | -     | -     | -     | -     | -     | 79.3   |
| Ours        | 98.7  | 86.9 | 93.6  | 56.0  | 61.1  | 69.6  | 76.9  | 80.5  | 93.8  | 71.6   | 95.9  | 87.3   | 71.8  | 96.0  | 75.3  | 87.1  | 79.3  | 70.6  | 77.7  | 80.5   |

Only finely-annotated categories are considered. Our approach achieves a mIoU score of 80.5% and obtains an absolute gain of 1.2%, compared to DFN [17]. Furthermore, our method outperforms existing methods in 18 out of 19 categories.

D. COMPARISON WITH OTHERS

Table 4 shows the comparison results on Cityscapes test set. Our method gets a comparable result by achieving a mIoU score of 80.5%. Furthermore, for some classes such as bus and pole the proposed method attains a significant performance improvement of 5% over existing approaches. Table 5 reports the results on ADE2K validation set. The baseline DenseASPP [10] obtains a mIoU score of 41.51%. Our approach provides superior results compared to existing methods with a mIoU of 44.0%.

In Figure 7 we present a comparison of segmentation maps generated by our network and the most popular methods [7], [9], [10] on ADE20K (first 3 rows) and Cityscapes (last 3 rows). It is apparent that our method produces results that are more precise, sharp and faithful to the ground-truth than those of the other competing approaches. For example, compare the detail in the following regions: row 1: side table, sofa cushion, window blinds and plant; row 2: table, screens and desktop PC; row 3: person and window; row 4: bus; row 5: fences and poles; row 6: traffic light, poles and persons in the background.

The proposed framework is not only efficient in segmenting large objects (see the accuracy of bus and train class in Table 4) but also the small objects. A visual illustration is presented in Figure 8, where it is evident that our method segments small objects with much higher accuracy than other competing methods. Note that only the cropped (zoomed-in) results are shown for better visualization.
FIGURE 7. Comparison with mainstream methods. Compare the following regions: window blinds and plant in row 1; table, and computer in row 2; person and light in row 3; bus in row 4; fences and poles in row 5 and row 6.

FIGURE 8. Performance on small objects. Row 1: input. Row 2: ground-truth. Row 3: PSPNet [7]. Row 4: DeepLabV3 [9]. Row 5: DenseASPP [10]. Row 6: Our results. Note that only the zoomed-in results (from the white bounding box in ground truth) are shown for better visualization.

E. IMPROVEMENTS ON OBJECT BOUNDARIES
In this section, we evaluate the segmentation accuracy with the trimap experiment to quantify the accuracy of different methods near object boundaries. Specifically, we first detect the boundaries of the labels on Cityscapes val set and apply the morphological dilation on these boundaries with void label annotations, which produces the dilated band (called trimap). We then compute the mean IoU for those remaining pixels. Fig. 9 shows the mean mIoU results of our method, DenseASPP [10] and ASPP [8]. Note that trimap width is equal to 0 means we do not have any operation on the label. We can see that there is a great gap between our method and other two methods without trimap. When we ignore some boundary information, the performance of DenseASPP [10] and ASPP [8] has greatly improved.

V. CONCLUSION
Existing semantic segmentation approaches generally put strong emphasis on high-level context aggregation, but do...
not explicitly include low- and mid-level information. Here, we show that an optimal combination of features present at all these three stages leads to a refined recovery of local details. Specifically, we propose a high-level context aggregation block, combining the strengths of dilated (sparse) and wider (dense) kernel, to enhance the expressive power of the network and avoid unfairly neglecting the local information. Besides, we employ a light-weight feature pyramid augmentation and semantic refinement unit to generate low- and mid-level features that are mixed with high-level features at the decoder. The designed modules are very flexible that can be easily migrated to any network. Our extensive experiments on two segmentation benchmarks show compelling improvements over the previous best techniques.

REFERENCES

[1] L. Fan, W. Wang, F. Zha, and J. Yan, “Exploring new backbone and attention module for semantic segmentation in street scenes,” IEEE Access, vol. 6, pp. 71566–71580, 2018.

[2] M. Wu, C. Zhang, J. Liu, L. Zhou, and X. Li, “Towards accurate high resolution satellite image semantic segmentation,” IEEE Access, vol. 7, pp. 55609–55619, 2019.

[3] Y. Yu, B. Jiang, and N. Lu, “A descriptor system approach for estimation of incipient faults with application to high-speed railway traction devices,” IEEE Trans. Syst. Man. Cybern., Syst., vol. 49, no. 10, pp. 2108–2118, Oct. 2019.

[4] L. Zhang, L. Zhang, B. Du, J. You, and D. Tao, “Hyperspectral image unsupervised classification by robust manifold matrix factorization,” Inf. Sci., vol. 485, pp. 154–169, Jun. 2019.

[5] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, “Feature learning using spatial-spectral hypervector discriminant analysis for hyperspectral image,” IEEE Trans. Cybern., vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[6] R. Dong, X. Pan, and F. Li, “DenseU-Net-based semantic segmentation of small objects in urban remote sensing images,” IEEE Access, vol. 7, pp. 65347–65356, 2019.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, Jul. 2017, pp. 2881–2890.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2017.

[9] L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587. [Online]. Available: https://arxiv.org/abs/1706.05587.

[10] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “DenseASPP for semantic segmentation in street scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 3684–3692.

[11] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., Munich, Germany, Oct. 2015, pp. 234–241.

[12] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters—improve semantic segmentation by global convolutional network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, Jul. 2017, pp. 4353–4361.

[13] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 801–818.

[14] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” 2018, arXiv:1805.10180. [Online]. Available: https://arxiv.org/abs/1805.10180.

[15] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, “FuseExc: Enhancing feature fusion for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 269–284.

[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., Honolulu, HI, USA, Jul. 2017, pp. 1925–1934.
[40] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, “ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 552–568.

[41] 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., Dec. 2015, pp. 1377–1385.

[42] Z. Wu, C. Shen, and A. Van Den Hengel, “Wider or deeper: Revisiting the resnet model for visual recognition,” 2016, arXiv:1611.10080 [Online]. Available: https://arxiv.org/abs/1611.10080

[43] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223.

[44] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ADE20K dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, Jul. 2017, pp. 633–641.

[45] T. Ke, J. Hwang, Z. Liu, and S. Yu, “Adaptive affinity fields for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Munich, Germany, Sep. 2018, pp. 587–602.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.

**LIYUAN LIU** received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2017, where she is currently pursuing the M.Sc. degree under the supervisor is Prof. Y. Pang. Her research interests include deep learning and its applications in scene parsing and object detection.

**YANWEI PANG** received the Ph.D. degree in electronic engineering from the University of Science and Technology of China, in 2004. He is currently a Professor with Tianjin University, China. His current research interests include object detection, image recognition, image processing, and their applications in self-driving cars and visual surveillance, and biometrics. He has published more than 120 scientific articles, including 37 IEEE Transactions articles in these areas.

**SYED WAQAS ZAMIR** received the Ph.D. degree in image processing from Universitat Pompeu Fabra, Spain, in 2017. He is currently working as a Senior Research Associate with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. In general, his research interests are in the field of image and video processing, color science, and computer vision.

**SALMAN KHAN** received the Ph.D. degree in computer vision from The University of Western Australia (UWA), under the supervision of W/Prof. M. Bennamoun. From 2016 to 2018, he served as a Research Affiliate with the Australian Center for Robotic Vision (ACRV). He is currently a Senior Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates.

**FAHAD SHAHBAZ KHAN** received the Ph.D. degree in computer vision from the Computer Vision Center Barcelona, Autonomous University of Barcelona, Spain. From 2014 to 2018, he was a Research Fellow with Linkoping University, Sweden. He is currently a Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates, and also an Associate Professor with the Computer Vision Laboratory, Linkoping University, Sweden.

**LING SHAO** received the Ph.D. (D.Phil.) degree in computer vision with the Robotics Research Group, University of Oxford, under the supervision of Prof. M. Brady. He was a Professor with the School of Computing Sciences, University of East Anglia, Norwich, U.K. He is currently the CEO and a Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.

---

**VOLUME 8, 2020**