Pyramid ALKNet for Semantic Parsing of Building Facade Image

Wenguang Ma, Wei Ma, *Member, IEEE*, Shibiao Xu, *Member, IEEE*, and Hongbin Zha, *Member, IEEE*

***Abstract*— The semantic parsing of building facade images is a fundamental yet challenging task in urban scene understanding. Existing works sought to tackle this task by using facade grammars or convolutional neural networks (CNNs). The former can hardly generate parsing results coherent with real images while the latter often fails to capture relationships among facade elements. In this letter, we propose a pyramid atrous large kernel (ALK) network (ALKNet) for the semantic segmentation of facade images. The pyramid ALKNet captures long-range dependencies among building elements by using ALK modules in multiscale feature maps. It makes full use of the regular structures of facades to aggregate useful nonlocal context information and thereby is capable of dealing with challenging image regions caused by occlusions, ambiguities, and so on. Experiments on both rectified and unrectified facade data sets show that ALKNet has better performances than those of state-of-the-art methods.**

***Index Terms*— Facade parsing, large kernel, man-made structure, nonlocal context.**

I. Introduction

FACADE parsing is the process of segmenting building facades into elements with semantic categories. It has a wide range of potential applications in various fields, including building reconstruction [2], [3], procedural modeling [4], urban planning [5], augmented reality/virtual reality (AR/VR), and city navigation systems [6], [7]. However, accurate facade parsing is very challenging due to the complexity of facade images. For example, parts of buildings are often occluded by trees or cars. Appearances of the same types of elements vary in facade data sets and even in single images due to shade, the opening-closing status of windows/balconies, and so on.

Considering that facades are man-made objects with regular structures, early approaches [8], [9] address the facade parsing problem by using shape grammars or shape priors which are able to generate neat segmentation results. However, these strong constraints generally ignore the variety of different facades, some parsing results are seriously inconsistent with the real buildings. Other methods [10]–[12] adopt hand-crafted features to segment facade images. In particular, Mathias *et al.* [11] proposed a three-layer structure that incorporates different levels of observations or principles. Gadde *et al.* [12] ignored domain-specific knowledge and used auto-context features to train classifiers for facade parsing. These methods can generate results that are more coherent with facades than grammar-based methods. However, they are not robust enough to tackle complex building scenes. Li *et al.* [13] introduced a systematic and hierarchical approach of parsing urban building facades from terrestrial-laser-scanning point clouds. Our method is designed for ordinary optical images that are easier to capture than point clouds.

Recent methods based on convolutional neural networks (CNNs), specifically being fully convolutional networks (FCNs [14]), have dominated the research of image segmentation [15]–[21]. For example, SegNet [15] stores the max-pooling indices, which are later used to upsample low-resolution features. U-Net [16] employs skip-connection to fuse low-level features to high-level ones, which performs well in biomedical image segmentation. More recently, Zhao *et al.* [19] proposed the pyramid scene parsing network (PSPNet), which adopts a multiscale feature fusion of pyramid pooling data to aggregate context from the sub-regions of different scales. Chen *et al.* [20] adopted an atrous spatial pyramid pooling module (PPM) to sparsely collect context information from surrounding pixels. These methods have demonstrated power in the pixel-wise classification of many types of scene images, for example, cars and pedestrians on roads, by integrating context around each pixel. The direct application of FCN methods on facade parsing leads to confusing results [1], [22]–[24] due to the challenges of facade parsing mentioned earlier.

To deal with facade parsing, we rethink the specialties of buildings: most elements, such as windows, balconies, and doors, have rectangular shapes in facade images; elements in each facade are aligned and distributed in a structural layout and the same type of elements from a facade is designed to look the same. Motivated by these observations, we propose a pyramid atrous large kernel (ALK) network (ALKNet) for facade parsing. The network is named by its core module, that is, pyramid ALK module, which is designed to aggregate rich nonlocal structural context information on multiple scales. Due to the multiscale structural context cues, pyramid ALKNet performs well with the challenging regions of facades.

Manuscript received November 30, 2019; revised March 29, 2020; accepted May 4, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61771026, Grant 61971418, and Grant 61671451 and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR). *(Corresponding author: Wei Ma.)*

Wenguang Ma and Wei Ma are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: mawenguang@emails.bjut.edu.cn; mawei@bjut.edu.cn).

Shibiao Xu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shibiao.xu@nlpr.ia.ac.cn).

Hongbin Zha is with the Key Laboratory of Machine Perception (MOE), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: zha@cis.pku.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.2993451

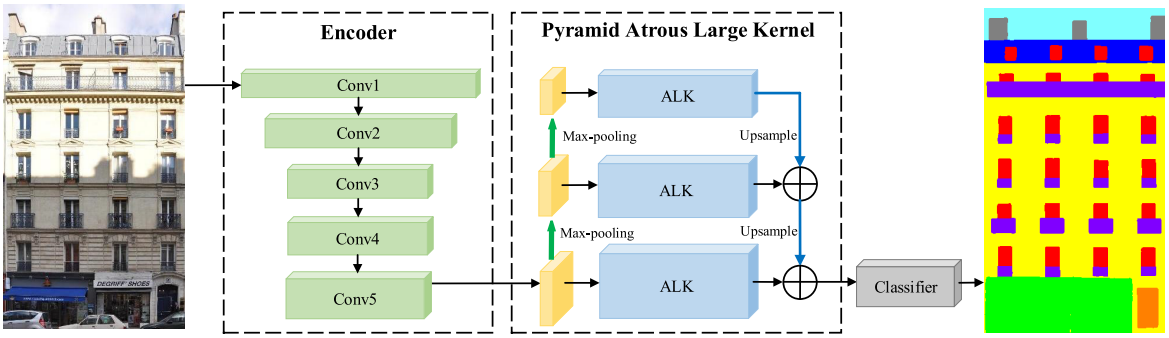

Fig. 1. Overview architecture of the proposed pyramid ALKNet. An encoder is adopted to extract features from the input facade image. The pyramid ALK module enhances the feature maps with structural context on multiple scales, before sending them to a classifier for facade parsing.

Our main contributions are threefold.

1) A novel pyramid ALKNet is proposed for facade parsing, which outperforms the state of the art in both rectified and unrectified facade data sets.
2) A pyramid ALK module is specially designed to encode nonlocal structures of facades, which is helpful in dealing with occlusions and appearance ambiguities.
3) Extensive comparisons and ablation experiments are conducted on two facade parsing benchmarks to validate the good performances of the proposed pyramid ALKNet and the key module of pyramid ALK.

II. Proposed Method

In this section, we explain the details of the proposed pyramid ALKNet. First, we present the overall network architecture. Then we provide the core part of the network, that is, the pyramid ALK module.

A. Our Network

The overview of the proposed pyramid ALKNet is shown in Fig. 1. It mainly includes an encoder and a pyramid ALK module. Given an input facade image, pyramid ALKNet first adopts ResNet-FCN as an encoder, which is designed in a fully convolutional manner to produce feature maps. To retain more details and produce more dense feature maps, we remove the downsampling operations in the last two layers and replace them with dilated convolutions as done in [17]. Thus, the output feature maps of the encoder become the eighth size of the input image.

After generating dense feature maps by the encoder, we use a 1×1 convolution to reduce the channel of "Conv5" features from 2048 to 512. Then, the feature maps are fed into the proposed pyramid ALK module to aggregate rich structural information. We build a feature pyramid architecture that applies an ALK module on each scale to capture the long-range dependencies among building elements. Finally, we apply a classifier, a 1×1 convolution layer, to perform pixel-wise classification on the enhanced features and another bilinear upsampling by a factor of 8 to generate the facade parsing maps.

B. Pyramid ALK Module

Current FCN-based methods are incapable of capturing the structural context of facades, especially those in complex scenes, for example, facades with occlusions caused by trees and cars, ambiguities due to shade, and so on. We note that most elements, such as windows, doors, and balconies, of building facade images have rectangular shapes. Furthermore, in every single image, the same type of elements, for example, windows, are aligned and look similar. Inspired by these characters, we design an ALK module to capture the nonlocal structural context.

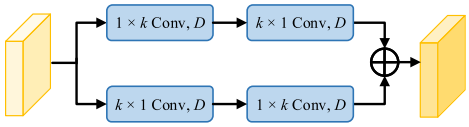

Fig. 2. ALK module. k and D are the kernel size and atrous rate, respectively.

As shown in Fig. 2, the ALK module contains two parallel branches. Each branch has two separable ALKs ($1 \times k$ and $k \times 1$, $k = 15$) with the same atrous rate D. Different from regular convolutions, atrous convolutions [17] sample the input features sparsely at first with intervals determined by the atrous rate D, to cover more spatial context without sacrificing computation cost. The atrous rate D varies with the size of the input features maps. On the other hand, large kernels can enlarge receptive fields and have been proven effective in classification and localization tasks [18]. We do not use large kernels directly since they are computationally complex. Additionally, considering the regular distribution of facade elements, we introduce separable large kernels. We perform the large kernel convolution vertically and then horizontally; the reverse order is also used. The two branches are different in their computation order for complementarity. The features enhanced by the two parallel branches are merged by element-wise addition.

To capture the long-range structural context of facades, instead of using a single ALK module, we adopt three ALK modules. Each ALK module takes a specific scale of feature maps as input. Compared with the one ALK module, this pyramid ALK module aggregates context more thoroughly. In implementation, as illustrated in Fig. 1, we first downsample the features generated by the encoder by 2×2 max-pooling with stride 2 and form a multiscale feature pyramid. For each scale of features, an ALK module with a specific atrous rate is applied for capturing the long-range structural context cues. In detail, the atrous rates $D = 4$, 2, and 1, for the three scales from bottom to top, respectively. After being enhanced by the ALK modules, the features of different scales are fused step by step

via element-wise addition, so as to incorporate the structural context of neighbor scales more precisely. Before each fusion, the coarser-resolution feature map is upsampled by a factor of 2 using bilinear upsampling. By enhancing the features with long-range and structural context, the pyramid ALK module successfully eliminates influence from occlusions and appearance ambiguities.

C. Loss Function

To train the pyramid ALKNet, we adopt a standard cross-entropy loss as follows:

$$\mathcal{L}_{ce} = \frac{1}{N}\sum_{i}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i)]$$

where N is the number of pixels, i is the index, and y and p are the ground truth probability distribution and predicted probability distribution of facade categories, respectively.

III. Experiments

To evaluate the proposed method, we carry out comprehensive experiments on the Ecole Centrale Paris (ECP) data set [25] and RueMonge2014 data set [26]. The experimental results demonstrate that our pyramid ALKNet achieves state-of-the-art performances.

A. Data Sets and Evaluation Metrics

1) ECP Data Set: The ECP data set [25] consists of 104 rectified facade images of Haussmannian style architectural buildings. The original annotations are imprecise, so we use the annotations provided by Mathias *et al.* [11], which consist of eight classes: window, wall, balcony, door, roof, chimney, sky, and shop. In our experiments, we follow the data splits of Liu *et al.* [1] and perform a fivefold cross-validation.

2) RueMonge2014: This data set, presented in [26], has 428 high-resolution and multiview images obtained from Paris streets. The images in this data set are annotated with seven classes: window, wall, balcony, door, roof, sky, and shop. Different from the ECP data set, the facade images in this data set are not rectified. We follow the training and test data splits as prespecified in [26].

3) Metrics: Our experiments adopt the mostly used metrics to quantitatively evaluate the performances of facade parsing, including the total accuracy (Total acc.), average class accuracy (Class avg.), F1 score, and mean intersection over union (IoU).

B. Implementation Details

We implement the proposed pyramid ALKNet based on TensorFlow and train it on a single GTX 1080Ti GPU. We adopt the ResNet50 [27] pretrained on ImageNet as the encoder, remove the last two downsampling operations and employ dilated convolutions in the subsequent convolutional layers. Adam [28] with a basic learning rate of 2e-4 and a weight decay of 0.0001 is used to optimize the whole network. To make the model robust, we adopt some data augmentation techniques, including random scaling, random cropping, and random horizontal flipping.

TABLE I

Quantitative Comparison on the ECP Data Set. The Best Performances for Each Metric Are Highlighted in Bold

Method	Total acc.	Class avg.	F1 score	Mean IoU
Kozinski *et al.* [9]	91.30	90.29	-	-
Yang *et al.* [10]	74.71	71.14	-	-
Mathias *et al.* [11]	88.02	85.22	-	-
Autocontext$_{2D}$ [12]	91.40	89.50	-	80.50
Cohen *et al.* [29]	90.34	88.63	-	-
DeepFacade [1]	93.54	90.57	91.03	83.78
Femiani *et al.* [23]	-	78.34	-	-
Fathalla *et al.* [24]	91.31	-	-	-
Rahmani *et al.* [30]	92.20	91.00	-	-
PSPNet [19]	93.17	90.82	91.06	83.78
ResNet50-FCN [27]	93.31	91.27	91.44	84.40
DeepLabv3+ [20]	93.35	91.20	91.37	84.29
DANet [21]	93.35	91.20	91.31	84.19
Ours	**93.56**	**91.67**	**91.74**	**84.90**

C. Comparisons on the ECP Data Set

The ECP data set [25] is widely used in existing facade parsing methods, which are compared with our method. Furthermore, since facade parsing belongs to a subfield of semantic segmentation, we also compare our method with state-of-the-art semantic segmentation methods. The experimental results on the ECP data set are shown in Table I, the first three rows are traditional facade parsing methods, CNN-based facade parsing methods and semantic segmentation methods, respectively.

Compared with the traditional methods, our method presents significant improvements because those hard constraints are inflexible for real facades, and hand-crafted features are not robust for complex building scenes. Among all these grammar-based methods, the best performances, 91.30% in total accuracy and 90.29% in class average accuracy, are achieved by [9]. Compared with this method, our model increases the total accuracy by 2.26% and the class average accuracy by 1.38%, respectively.

Our method also substantially outperforms CNN-based facade parsing methods [23], [24] which are limited to local receptive fields and short-range context. DeepFacade [1] is the most competitive method. It used a symmetrical loss function based on the symmetry of facade elements, which is combined with a cross-entropy loss function to optimize parsing results. Then, a refinement process is performed based on detected bounding boxes. Considering that the refinement process is empirical and highly data set-dependent, and cannot be applied to unrectified facades, in agreement with Liu *et al.* [1], we compare the outputs of segmentation networks rather than those after being postprocessed. From Table I, it can be seen that our method outperforms DeepFacade [1] in all metrics.

For comparisons with state-of-the-art semantic segmentation methods, we use the same experimental settings and hyperparameters to train them on the ECP data set. The results are shown in the third row of Table I. PSPNet [19] and Deeplabv3+ [20] adopt PPM and atrous spatial pyramid pooling (ASPP), respectively, to collect contextual information. Compared with these approaches, the pyramid ALKNet achieves superior performances because our method captures the context information by aggregating the long-range

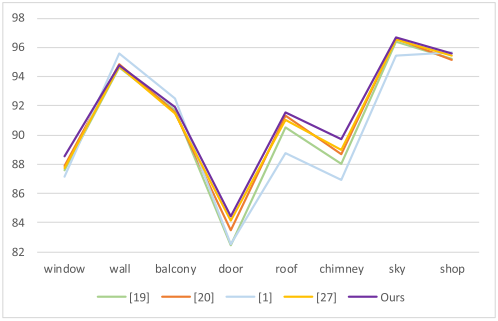

Fig. 3. Class accuracies of CNN-based methods on the ECP data set.

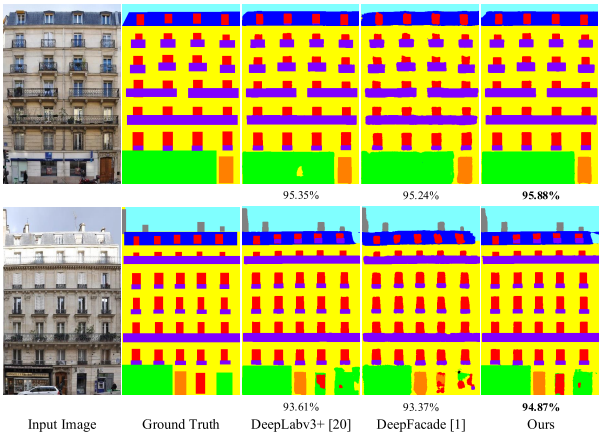

Fig. 4. Visual comparison on the ECP data set.

dependencies from horizontal and vertical directions which is consistent with the structures of facades.

We choose the CNN-based models, PSPNet [19], DeepLabv3+ [20], DeepFacade [1], and ResNet50-FCN [27], which perform the best among all the methods listed in Table I, and we present their accuracies on each class in Fig. 3. From the figure, it can be seen that our model has a leading performance in most classes. Our accuracies on the wall and balcony classes are slightly lower than those of DeepFacade [1]. This is because our model sacrifices a certain accuracy on walls and balconies to keep the shape and distribution regularity of elements such as windows, which is essential for downstream applications, for example, 3-D reconstruction.

Fig. 4 shows the visual results obtained on the ECP data set. Compared with DeepFacade [1], which adopts a symmetric loss to restraint the shape of elements, our model obtains more rectangular and regularly distributed results due to the specially designed ALK module that captures the structural context of facades well. Additionally, compared with typical contextual information aggregation modules, such as PPM [19] and ASPP [20], the structural context captured by our method is more discriminative and capable of dealing with challenging image regions caused by occlusions and ambiguities.

D. Comparisons on the RueMonge2014 Data Set

We here compare our method with state-of-the-art methods on the RueMonge2014 data set [26]. Different from the ECP data set in which all images have been rectified, the images in RueMonge2014 are raw and tilted (see Fig. 5). The experimental results are given in Table II.

TABLE II
QUANTITATIVE COMPARISON ON THE RUEMONGE2014 DATA SET

Method	Total acc.	Class avg.	F1 score	Mean IoU
Autocontext$_{2D}$ [12]	81.20	73.70	-	60.50
Autocontext$_{2D-3D}$ [12]	81.90	79.00	-	62.70
SPLATNet$_{2D}$ [31]	-	-	-	69.30
SPLATNet$_{2D-3D}$ [31]	-	-	-	70.60
PSPNet [19]	87.24	80.60	82.12	70.93
ResNet50-FCN [27]	88.18	82.54	83.61	72.85
DeepLabv3+ [20]	88.09	82.65	83.73	72.96
DANet [21]	87.86	82.40	83.53	72.69
Ours	**88.40**	**83.86**	**84.09**	**73.42**

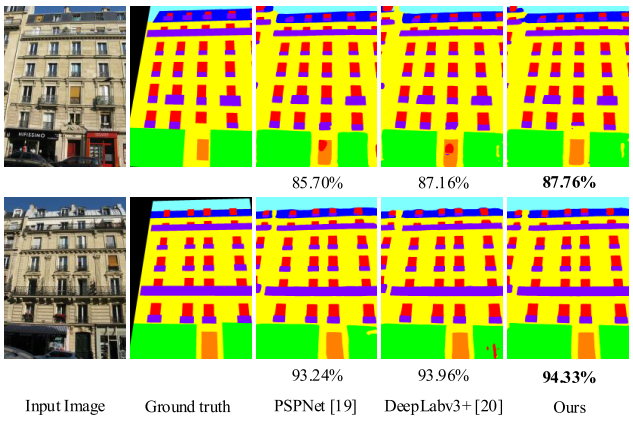

Fig. 5. Visual comparison on the RueMonge2014 data set.

Compared with state-of-the-art semantic segmentation methods, such as DeepLabv3+ [20], our model achieves better performances in all metrics. This result is mainly because the pyramid ALK module explores structural context to help eliminate ambiguities in challenging regions. Note that facade elements in these tiled images are not as regularly distributed as those in the rectified pictures. However, they are still distributed horizontally and approximately vertically, especially in small-scale views. Therefore, the pyramid ALK module can still aggregate structural context from these tiled images. Autocontext$_{2D\text{-}3D}$ [12] and SPLATNet$_{2D\text{-}3D}$ [31] use point cloud data for joint 2D–3D processing, but our method only uses inputs of 2D images. Regardless, our method outperforms these significantly. Fig. 5 shows certain facade parsing results on the RueMonge2014 data set. Compared with PSPNet [19] and DeepLabv3+ [20], which are popular in semantic segmentation tasks, our method overcomes the challenges of unrectified facades and obtains better visual qualities.

E. Ablation Study

To verify the settings in the proposed pyramid ALK module, we carry out ablation experiments on the ECP data set. For a fair comparison, we change only one setting in each ablation study and keep the others at their optimum settings. First, we attempt to use different layers in the Pyramid ALK module. The settings "{1}," "{2}," and "{1, 2, 3}" in the first row of Table III indicate the pyramid layers used in the decoder of our model. Then, we try different kernel sizes ranging from 11 to 15. The max kernel size is 15 due to the minimum feature map size of 16 $\times$ 16. At last, we attempt to fuse features from different levels of the encoder. By using a 1 $\times$ 1 convolution to adjust the channel of each feature to 512, these features are then fused by element-wise addition.

TABLE III
ABLATION STUDIES OF THE PYRAMID ALK ON THE ECP DATA SET

Setting	Total acc.	Class avg.	F1 score	Mean IoU
{1}	93.47	91.55	91.59	84.65
{1, 2}	93.50	91.56	91.65	84.74
{1, 2, 3}	**93.56**	**91.67**	**91.74**	**84.90**
11	93.48	91.55	91.69	84.81
13	93.44	91.52	91.59	84.64
15	**93.56**	**91.67**	**91.74**	**84.90**
Conv3, 4, 5	93.39	91.37	91.46	84.45
Conv4, 5	93.39	91.46	91.58	84.62
Conv5	**93.56**	**91.67**	**91.74**	**84.90**

Table III shows the parsing results of ALKNet with different pyramid layers, kernel sizes, and encoder features. From the table, it can be seen that our model obtains the best results when using three pyramid layers ("{1, 2, 3}"). With the number of pyramid layers increasing, the structural context becomes richer and all metrics are improved. Moreover, the pyramid ALK module benefits the most when the kernel size is 15. The highest-level features ("Conv5") are the most useful for facade parsing. Adding low-level features that contain excessively detailed information would slightly decrease the performance.

IV. CONCLUSION

In this letter, pyramid ALKNet was presented for the semantic parsing of building facade images. The core of the model is a pyramid ALK module, specially designed to make full use of the structure of facades to extract discriminative features by encoding nonlocal structural context information. Due to the pyramid ALK module, our ALKNet is capable of dealing with challenging image regions caused by occlusions, ambiguities, and so on. Extensive comparison experiments on rectified and unrectified facade data sets both demonstrate that the proposed pyramid ALKNet exhibits outstanding performances.

ACKNOWLEDGMENT

We would like to thank the authors of DeepFacade [1] for providing us their facade parsing results and constructive suggestions on comparative experiments.

REFERENCES

[1] H. Liu, J. Zhang, J. Zhu, and S. C. H. Hoi, "DeepFacade: A deep learning approach to facade parsing," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2301–2307.

[2] M. Shahzad and X. Xiang Zhu, "Robust reconstruction of building facades for large areas using spaceborne TomoSAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 752–769, Feb. 2015.

[3] B. Wu, X. Sun, Q. Wu, M. Yan, H. Wang, and K. Fu, "Building reconstruction from high-resolution multiview aerial imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 855–859, Apr. 2015.

[4] B. Hohmann, S. Havemann, U. Krispel, and D. Fellner, "A GML shape grammar for semantically enriched 3D building models," *Comput. Graph.*, vol. 34, no. 4, pp. 322–334, Aug. 2010.

[5] D. Gonzalez-Aguilera, E. Crespo-Matellan, D. Hernandez-Lopez, and P. Rodriguez-Gonzalvez, "Automated urban analysis based on LiDAR-derived building models," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1844–1851, Mar. 2013.

[6] W. Ma and W. Ma, "Deep window detection in street scenes," *KSII Trans. Internet Inf. Syst.*, vol. 14, no. 2, pp. 855–870, Feb. 2020.

[7] K. Lee, Y. Kim, S. I. Cho, and K. Choi, "Building detection in augmented reality based navigation system," in *Proc. 13th Int. Conf. Multimedia Modeling*, vol. 2, Jan. 2006, pp. 544–551.

[8] O. Teboul, L. Simon, P. Koutsourakis, and N. Paragios, "Segmentation of building facades using procedural shape priors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3105–3112.

[9] M. Kozinski, R. Gadde, S. Zagoruyko, G. Obozinski, and R. Marlet, "A MRF shape prior for facade parsing with occlusions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2820–2828.

[10] M. Y. Yang and W. Förstner, "Regionwise classification of building facade images," in *Proc. ISPRS Conf. Photogramm. Image Anal.*, 2011, pp. 209–220.

[11] M. Mathias, A. Martinović, and L. Van Gool, "ATLAS: A three-layered approach to facade parsing," *Int. J. Comput. Vis.*, vol. 118, no. 1, pp. 22–48, May 2016.

[12] R. Gadde, V. Jampani, R. Marlet, and P. V. Gehler, "Efficient 2D and 3D facade segmentation using auto-context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1273–1280, May 2018.

[13] Z. Li *et al.*, "A hierarchical methodology for urban facade parsing from TLS point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 123, pp. 75–93, Jan. 2017.

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

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[16] 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*, 2015, pp. 234–241.

[17] 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. 2018.

[18] 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. (CVPR)*, Jul. 2017, pp. 4353–4361.

[19] 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.

[20] L.-C. 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.*, Sep. 2018, pp. 801–818.

[21] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[22] M. Schmitz and H. Mayer, "A convolutional network for semantic facade segmentation and interpretation," *ISPRS Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. XLI–B3, pp. 709–715, Jun. 2016.

[23] J. Femiani, W. Reyaz Para, N. Mitra, and P. Wonka, "Facade segmentation in the wild," 2018, *arXiv:1805.08634*. [Online]. Available: http://arxiv.org/abs/1805.08634

[24] R. Fathalla and G. Vogiatzis, "A deep learning pipeline for semantic facade segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 120.1–120.13.

[25] O. Teboul. (2010). *Ecole Centrale Paris Facades Database*. Accessed: Nov. 1, 2019. [Online]. Available: http://vision.mas.ecp.fr/Personnel/teboul/data.php

[26] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool, "Learning where to classify in multi-view semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 516–532.

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

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[29] A. Cohen, A. G. Schwing, and M. Pollefeys, "Efficient structured parsing of facades using dynamic programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3206–3213.

[30] K. Rahmani and H. Mayer, "High quality facade segmentation based on structured random forest, region proposal network and rectangular fitting," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. IV–2, pp. 223–230, May 2018.

[31] H. Su *et al.*, "SPLATNet: Sparse lattice networks for point cloud processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2530–2539.