Copyright ⓒ 2022 KSII

Deep Facade Parsing with Occlusions

Wenguang Ma[1], Wei Ma[1*], and Shibiao Xu[2]
[1] Faculty of Information Technology, Beijing University of Technology
Beijing 100124, China
[e-mail: mawenguang@emails.bjut.edu.cn, mawei@bjut.edu.cn]
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
Beijing 100190, China
[e-mail: shibiao.xu@nlpr.ia.ac.cn]
*Corresponding author: Wei Ma

Received January 25, 2021; revised August 31, 2021; revised November 17, 2021; accepted January 27, 2022; published February 28, 2022

Abstract

Correct facade image parsing is essential to the semantic understanding of outdoor scenes. Unfortunately, there are often various occlusions in front of buildings, which fails many existing methods. In this paper, we propose an end-to-end deep network for facade parsing with occlusions. The network learns to decompose an input image into visible and invisible parts by occlusion reasoning. Then, a context aggregation module is proposed to collect nonlocal cues for semantic segmentation of the visible part. In addition, considering the regularity of man-made buildings, a repetitive pattern completion branch is designed to infer the contents in the invisible regions by referring to the visible part. Finally, the parsing map of the input facade image is generated by fusing the results of the visible and invisible results. Experiments on both synthetic and real datasets demonstrate that the proposed method outperforms state-of-the-art methods in parsing facades with occlusions. Moreover, we applied our method in applications of image inpainting and 3D semantic modeling.

Keywords: Facade parsing, occlusion, repetitive pattern, man-made structure

This work was supported in part by the National Natural Science Foundation of China under Grant 62176010, Grant 61771026, and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

http://doi.org/10.3837/tiis.2022.02.009 **ISSN : 1976-7277**

1. Introduction

The purpose of facade parsing is to segment rectified facade images into semantic elements, including windows, doors, and balconies, etc. It is a key step in urban scene understanding and can also help other tasks. Unfortunately, serious occlusions are quite common in facade images, as shown in **Fig. 1**. How to parse buildings with occlusions is an inevitable but challenging problem. Addressing the problem will bring great benefits to many applications, e.g., city surveying and mapping, completion of facade images, and semantic 3D reconstruction of buildings.

Traditional methods for facade parsing [1] attempt to address the occlusion problem by using grammars or priors predefined for man-made buildings. It is difficult for these methods to generate results coherent with specific input images. In contrast, facade parsing is the semantic segmentation of facade images. Recently, many CNN-based models for semantic segmentation of indoor scenes [2] and street scenes [3], [4] have been developed and have shown performances beyond traditional methods. These models have also been adapted to parse facade images, e.g., by integrating facade-specific losses [5]. However, these models are designed to learn for pixel-wise classification according to the appearances around the pixels. Directly training them to predict labels for invisible pixels will confuse them in terms of the attributes of facade elements. The results obtained in this way, e.g., by using the model of DeepFacade [5], are generally cluttered in occluded regions, as shown in **Fig. 1**.

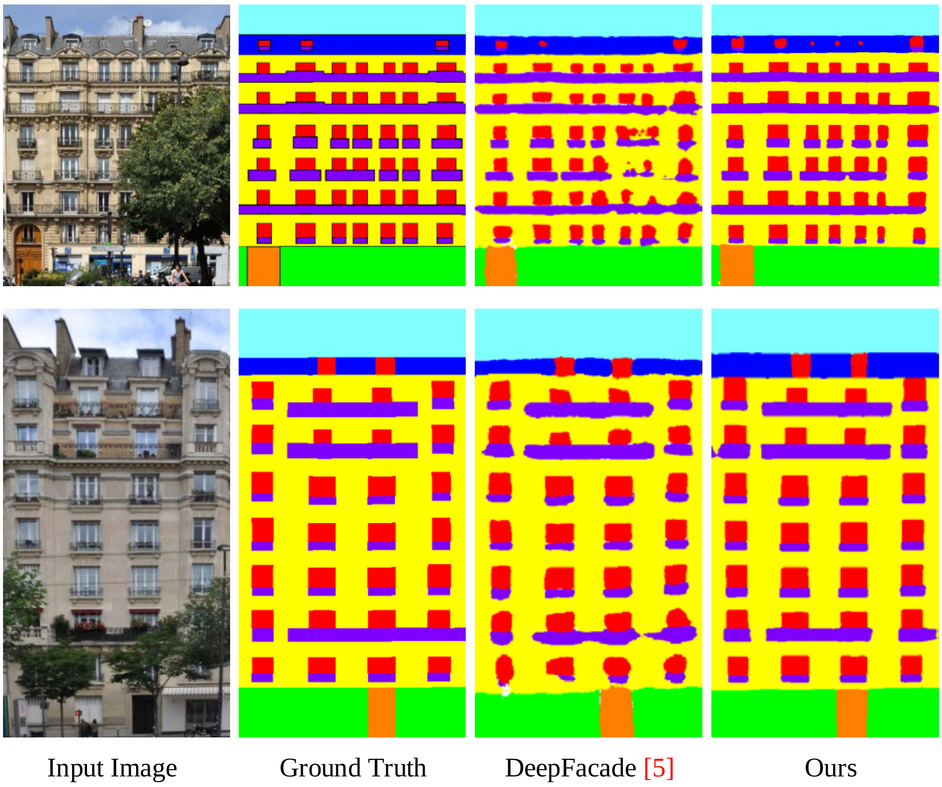


Fig. 1. Facade parsing results obtained by DeepFacade [5] and the proposed method.

Repetitive patterns, such as windows, balconies and tiles, are the most prominent and significant features in urban buildings. Existing facade parsing methods [6], [7] adopt a strong repetition assumption for analyzing facade structures. These hard-coded methods cannot deal with occlusion if the parsed buildings violate prior assumptions. Different from them, we model the repetitive patterns of building facades as prior knowledge into deep neural networks and learn to reason the contents of occluded regions with the help of visible parts, which is of great importance to the facade parsing with occlusions.

In this paper, we present an end-to-end deep network for facade parsing with occlusions. It learns to decompose an input image into visible and invisible parts via occlusion reasoning. For visible parts, a specifically designed segmentation branch capable of aggregating nonlocal context is adopted to infer their labels. To further refine the results obtained by the context aggregation module, we leverage the regularity in man-made buildings as done in traditional methods. In contrast, we design a repetitive pattern completion network, which can learn to infer contents in occluded parts by visible parts in the same image. Compared with those generated by predefined priors/grammars, our results are more realistic. Finally, we merge the results in the visible and invisible parts with a fusion module. We perform an evaluation on two facade datasets with different degrees of occlusions: ECP-occluded and ENPC Art-deco [8]. ENPC Art-deco is a public benchmark dataset, in which most of the images have relatively small occlusions (occupying less than 18% of an image). ECP-occluded is a synthesized dataset generated by adding collected trees to the original clean ECP data [9] to simulate facade images captured along streets with luxuriant trees.

Our main contributions are four-fold:

- We propose a novel deep architecture for end-to-end facade parsing, which is the first work to effectively deal with facade parsing with large occlusions in the framework of deep learning.
- We design a context aggregation module to capture nonlocal context information of facades and propose a repetitive pattern completion network (RPCNet) by considering the regularity of man-made buildings and the cues of visible parts to infer contents in occluded regions.
- We apply joint multi-task learning and perform extensive comparisons and ablation experiments to verify the proposed architecture and its key components.
- To further evaluate the effectiveness of the proposed method, we provide the application implementations of image inpainting and 3D semantic modeling, which obtain promising results.

2. Related Work

2.1 Traditional methods

Traditional methods are highly dependent on hand-crafted knowledge priors. Several methods [1], [10]–[12] have attempted to solve facade parsing with occlusions by using shape grammars or hard constraints defined for buildings. For example, [1] adopted user-defined shape priors that encode the regular structures and layouts of elements to semantically decompose occluded regions. [11] proposed parsing facade images using dynamic programming with hard constraints. Inspired by them, [12] dealt with occlusions by leveraging the symmetry and repetitions of building elements. These rule-based traditional methods have difficulty generating realistic parsing results coherent with input images. Additionally, some works [13] typically learned to discriminate different elements by using handcrafted features and pixel-wise classifiers. [14] adopted the auto-context features for 2D and 3D facade segmentation. In general, both of these methods are incapable of handling occlusions.

The repetitive pattern is an important attribute for facade parsing and occluded content inference. The early method [15] introduced an approach for separating and segmenting individual facades using prior knowledge of repetitive patterns. [6] combined low-level classifiers with mid-level object detectors to infer an irregular lattice for facade parsing. [7] combined traditional feature extraction processes and a Kronecker product low-rank model to detect the repetitive patterns of facade structures. The method can only parse the specific repetitive pattern, such as windows, rather than all facade labels. It would fail if the Kronecker product model cannot represent the structures of facades. However, our method takes full advantage of the deep learning techniques and learns the repetitive patterns from data. [16] proposed a pipeline to extract and synthesize repetitive patterns in single images. We mainly focus on using the repeatability of buildings to improve the facade parsing results.

2.2 Deep learning-based methods

Compared to traditional methods, deep learning frameworks, specifically being fully convolutional networks (FCNs) [17], were proved to be powerful for pixel-wise classification. For example, U-Net [18] employed skip-connections to fuse the low-level features to high-level ones. The low-level features mainly include edges and textures of objects and the high-level ones include some semantic information. SegNet [19] stored the index of max-pooling which is later used to upsampling low-resolution features. However, these methods only perform pixel-wise classification well in visible parts. They cannot handle inference in invisible regions. The surrounding context is known to be helpful in dealing with occlusions and many context aggregation methods have appeared [20], [21]. For example, PSPNet [20] proposed a pyramid pooling module to aggregate context from sub-regions at multiple scales. DeepLabv3+ [21] used dilated convolutions of different rates for context information aggregation. However, context information is competent just for small occlusions. It is quite challenging to handle serious occlusions. Instead of only adopting a context aggregation module, we specially design a new branch incorporating the regularity in man-made buildings.

There are also some deep learning-based methods that are proposed for facade parsing. [22] proposed a fully convolutional network to parse some components cropped from facades. [23] applied FCNs to obtain the most likely label of each pixel, then the results were optimized through Restricted Boltzmann Machines by adopting horizontal and vertical scanlines. [24] proposed three networks to achieve multilabel semantic segmentation results of facade images. DeepFacade [5], [25] adopted a symmetric regularization which includes a rectangle constraint and a detector constraint to incorporate the shape knowledge of facade elements. Inspired by the regularities of building facades, Pyramid ALKNet [26] proposed the pyramid atrous large kernel module to extract the nonlocal structural context information. [27] presented a multiview architecture to collect reliable and visible clues from nearby views and used these clues to enhance the feature representation of a target view. Some other works mainly focus on window detection [28], [29] and building facade reconstruction [30]. The above methods usually require clean building facades. As far as we know, there is no work dealing with facade parsing with serious occlusions in the framework of deep learning.

3. Proposed Architecture

In this section, we first present the end-to-end deep facade parsing architecture to address occlusions, which is shown in **Fig. 2**. Next, we explain details of the coarse semantic segmentation branch and the repetitive pattern completion branch which are designed to infer contents in occluded regions. Finally, the fusion module and the loss function for training each branch are provided.

Our method (**Fig. 2**) is based on the encoder architecture. The encoder pre-trained on ImageNet extracts features that are an eighth of the size of the input image. We remove the last two downsampling operations of the encoder and employ the dilated convolutions to obtain more details and produce dense features. The common features extracted from the encoder are fed into a **coarse semantic segmentation** branch and a **repetitive pattern completion** branch. The coarse semantic segmentation branch captures the nonlocal structural context information for parsing the visible part of the building facade correctly. The repetitive pattern completion branch first reasons for the main occlusions of the input image, and then infers the contents of occluded regions using the coarse facade parsing results. Finally, the coarse semantic segmentation map and the completed repetitive patterns are fed into the fusion module to produce the final facade parsing result.

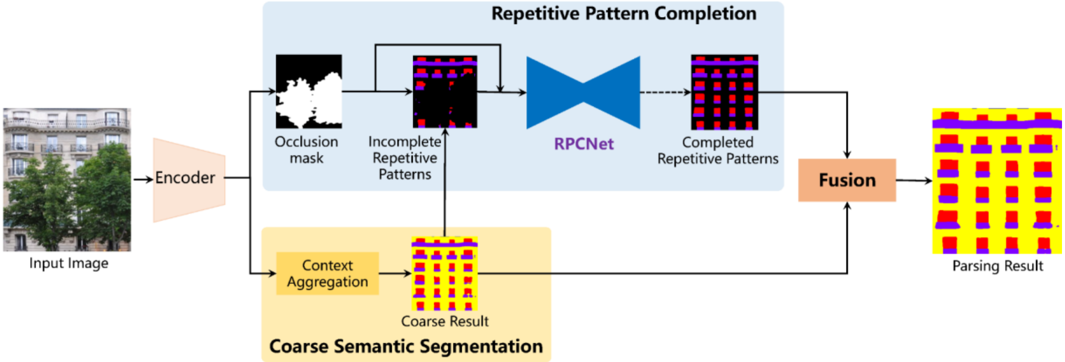

Repetitive Pattern Completion
Occlusion mask
Incomplete Repetitive Patterns
RPCNet
Completed Repetitive Patterns
Input Image
Encoder
Fusion
Parsing Result
Context Aggregation
Coarse Result
Coarse Semantic Segmentation

Fig. 2. Overview of the proposed architecture. An encoder is adopted to extract common features from the input image. The coarse semantic segmentation branch employs a context aggregation module to perform pixel-wise classification on the whole image. The repetitive pattern completion branch first produces an occlusion mask using common features. Then, elements of repetitive patterns in the visible regions are fed into RPCNet. RPCNet reasons the contents in the occluded parts. Finally, completed repetitive patterns are fused with the coarse semantic segmentation result.

3.1 Coarse Semantic Segmentation

We expect CNN to be able to capture the man-made structures of building facades. However, current FCN-based methods are incapable of aggregating the structural context of facades due to the complex scenes, such as the occlusions caused by trees and cars. Most elements, such as windows, doors, and balconies, of building facades have rectangular shapes. Typically, in every single image, these elements are aligned and similar. Inspired by these characteristics, we formulate the intrinsic structure information of facades into deep neural networks and propose the context aggregation module. The context aggregation module obtains the structures of facade images by capturing the nonlocal context along with the horizontal and vertical directions from multiple effective fields-of-view.

With the common features extracted by the encoder, the coarse semantic segmentation branch adopts a context aggregation module to capture nonlocal context information of facades. In detail, we employ the original atrous spatial pyramid pooling (ASPP) [31] (see **Fig. 3**) which consists of one 1 × 1 convolution and three 3 × 3 convolutions with rate=(12, 24, 36), and image-level features (all with 256 channels). The ASPP module collects multi-scale context information only from a few surrounding pixels and cannot capture dense semantics actually. Therefore, we add a residual large kernel (see **Fig. 3**) after the ASPP module. The residual large kernel contains a parallel 1 × k and k × 1 convolutions and a shortcut connection for performing identity mapping. Their outputs are simply fused by elementwise addition. The k is fixed at 15 in our experiments. The two 1D large kernels are efficient in producing discriminative features because facade elements, such as windows and balconies, are typically aligned horizontally and vertically. Thus, the common features from the encoder are greatly enhanced by our context aggregation module. A classifier uses the contextual features to generate a coarse facade parsing result of the whole image.

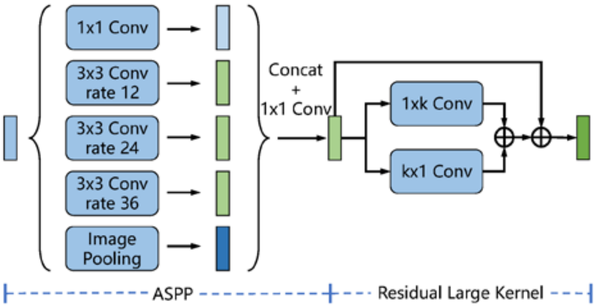


Fig. 3. The proposed context aggregation module. k means the kernel size in the 1D convolution layer.

3.2 Repetitive Pattern Completion

There are some challenging occluded repetitive patterns that are quite common in facade images, such as nonaligned windows, elements of different widths and complex occlusion objects. Particularly, the nonaligned windows are inconsistent with some hard rules of man-made structures, which straightforwardly causes grammar-based methods to generate incorrect parsing results. In addition, the width differences in surrounding windows and the gap differences between elements lead to serious irregular lattices. Constraint-based traditional methods cannot handle these diverse and complex patterns. Furthermore, occlusions vary in natural scenes. For example, the contours and sizes of occlusions are not fixed, and the categories of occluded objects are unknown. The densities of occlusions that greatly affect the parsing results are always uncertain. The appearances in the occluded region can be perceived through sparse occlusions, but cannot be perceived under dense occlusions.

To further refine the facade parsing results, we leverage the repetitions in man-made buildings. In particular, we design a repetitive pattern completion branch to infer the contents of occluded regions depending on the repetitive patterns of visible regions. For the occluded part, we employ a residual block [32] to transfer and reduce the dimension of common features. The occlusion mask is then generated by a classifier. Since windows and balconies are the most repetitive elements in facades, we use the two-class layers after the softmax activation from the coarse semantic segmentation branch. Again, the occlusion mask after softmax activation with window and balcony layers is used to form the incomplete repetitive patterns (see **Fig. 2**). The occlusion mask and incomplete repetitive patterns are concatenated and sent to the repetitive pattern completion network together.

3.2.1 Repetitive Pattern Completion Network

The repetitive pattern completion network, abbreviated as RPCNet, uses the regularities of repetitive patterns in visible parts to infer the contents in occluded parts. The most repetitive pattern in the facade is the window-balcony combination, which always has a rectangular shape; balconies are required to be located below windows, and almost all pairs of window-balconies lie on the same floor and have the same height. Therefore, we take the combination of windows and balconies (see **Fig. 4**) as a repetitive pattern of facades. The RPCNet is designed to complete the incomplete repetitive patterns using these characteristics of rectangular shapes and repetitive patterns of man-made structures in visible regions. Moreover, in order to achieve accurate completion results, the repetitive pattern completion branch recurrently adopts the RPCNet to infer the contents in occluded regions.

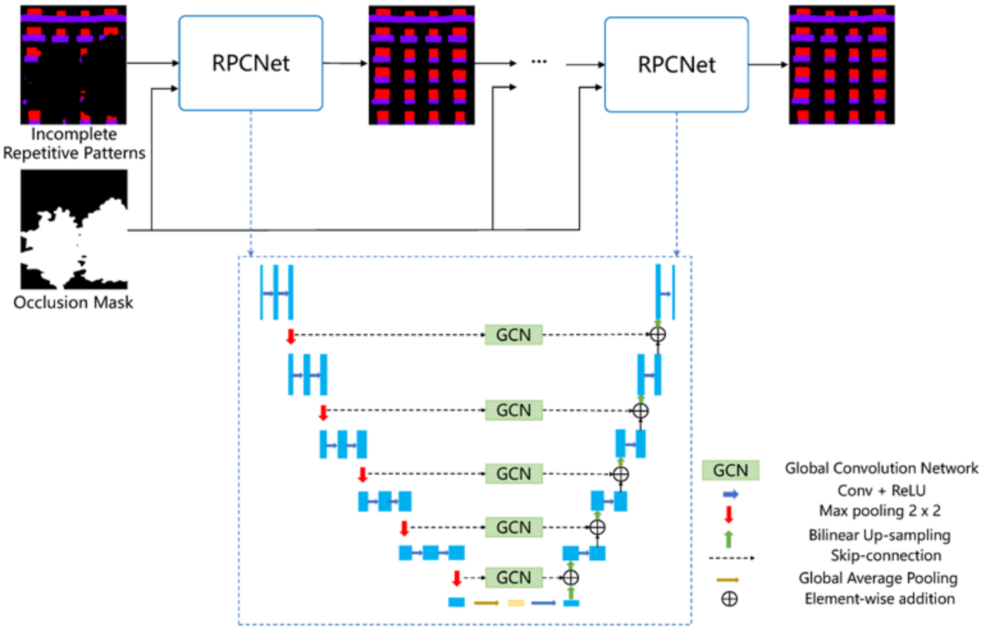


Fig. 4. The overview of the repetitive pattern completion network. The incomplete repetitive patterns and occlusion mask are iteratively fed into RPCNet to infer the contents in occluded regions.

As shown in **Fig. 4**, the incomplete repetitive patterns and the occlusion mask are concatenated to form the inputs of RPCNet. The occlusion mask provides guidance for RPCNet to infer the contents in occluded regions. RPCNet is a U-shaped network that has been proven to be useful in image completion [33] and image-to-image translation [34]. To capture the whole context information of the input image, we use image-level features, like [35]. Global average pooling is applied on the last feature map of the encoder, followed by a 1 × 1 convolution to adjust the channel and then bilinear upsampling is applied for the features. Next, a global convolutional network (GCN) [36] which consists of a combination of 1 × k + k × 1 and k × 1 + 1 × k convolutions, is employed to aggregate the global context information of each scale (see **Fig. 4**). In our experiments, k is fixed at 15. RPCNet adopts a top-down fusion strategy, and we first upsample the spatial resolution of a coarse resolution feature map by a factor of 2 through bilinear upsampling. Then the upsampled map is fused with the corresponding bottom-up feature map enhanced by GCN via element-wise addition. Finally, RPCNet generates a parsing result that has the same resolution and channel with the input repetitive patterns. The parsing result demonstrates that the context information for shapes and layouts from multi-scale features are well captured by RPCNet.

In particular, the repetitive patterns completed by a previous RPCNet and the occlusion mask are concatenated again, and fed into the next RPCNet to obtain more accurate repetitive pattern results. The recurrent completion process improves the quality of occluded patterns, as shown in **Fig. 5**. With the increasing number of iterations, the shape of each pattern and the layouts of facades become better. In the implementation, we experimentally set the number of iterations to 5. Note that the parameters of all RPCNet are shared.

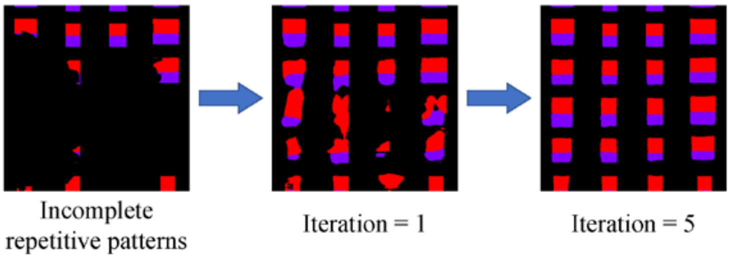


Fig. 5. Completion results of RPCNet with different iterations.

3.2.2 Synthetic Repetitive Patterns

Considering that there are too few training facade images and enhancing the completion results, we synthesize some repetitive pattern images with occlusions for training the proposed RPCNet. The synthetic training images which contain window and balcony combination. Specifically, we first generate images and repetitive patterns of random heights and widths. Then, we adopt an occlusion mask of random size to cover repetitive patterns at random locations. In the experiment, we first generate 10000 synthetic images for pretraining the RPCNet and use the facade training set to finetune the RPCNet. RPCNet is not only an end-to-end deep network that can be used as plug-and-play in any state-of-the-art semantic segmentation method, but can also handle quite challenging occlusions. For example, in **Fig. 5**, even though the occlusion region is large and the sizes of windows are different, our RPCNet can handle these serious occlusions by learning the diverse repetitive patterns from data.

3.3 Fusion Module

In **Fig. 6**, we design a fusion module by considering the repetitive and non-repetitive regions of facades. For example, windows are generally grid-like, but some balconies cross more than one window or even the whole facade horizontally. The proposed fusion module can handle this irregular phenomenon. After obtaining the completed repetitive patterns $P_{completed}$ from the repetitive pattern completion branch, we feed these patterns into a fusion module together with the coarse semantic segmentation result S_{coarse}. First, we produce the repetitive mask $M_{repetitive}$ and non-repetitive mask $M_{non-repetitive}$ according to the $P_{completed}$. The window label and balcony label in $P_{completed}$ are marked as repetitive mask $M_{repetitive}$, and the rest of the pixels are marked as non-repetitive mask $M_{non-repetitive}$. Then, we crop the repetitive patterns in $M_{repetitive}$ and coarse results in $M_{non-repetitive}$. After cropping, we fuse the repetitive patterns and the coarse result to generate a final facade parsing result S_{final}. Therefore, our fusion module can not only handle the irregular layout of balconies through the powerful context aggregation module of the coarse semantic segmentation branch, but also effectively infers the contents in occluded regions through the repetitive pattern completion branch.

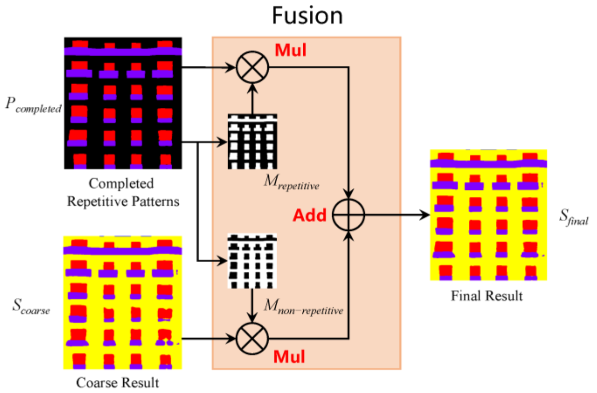


Fig. 6. The pipeline of the fusion module. The completed repetitive patterns and the coarse results are from the repetitive pattern completion branch and coarse semantic segmentation branch, respectively.

3.4 Joint Multi-task Learning

To train the overall networks, we jointly learn the coarse semantic segmentation, occlusion reasoning and repetitive pattern completion together. Here, standard cross-entropy (CE) is used on predicted coarse semantic segmentation s, occlusion mask o and repetitive patterns p. The final loss function can be written as:

$$L = L_{coarse}(s, \hat{s}) + L_{occlusion}(o, \hat{o}) + L_{pattern}(p, \hat{p}) \tag{1}$$

where $\hat{s}$ is the ground truth probability distribution of semantic labels, $\hat{o}$ is the ground truth probability of the occlusion mask and $\hat{p}$ is the ground truth probability distribution of repetitive patterns, respectively. As depicted in **Fig. 2**, the CE supervision on coarse semantic segmentation, occlusion mask and repetitive patterns are performed before feeding them into the fusion module.

4. Experiments

4.1 Datasets and Metrics

To evaluate the effectiveness of the proposed method, we conduct comprehensive experiments on the ECP-occluded dataset and the ENPC Art-deco dataset.

The ECP dataset [9] has 104 rectified facade images of Haussmannian style buildings that involve eight classes: window, wall, balcony, door, roof, chimney, sky, and shop. The original annotations are imprecise, so we use the annotations reannotated by [13]. In our experiments, we perform 5-fold cross-validation on this dataset. Since ECP is a clean facade dataset that has no significant occlusions, we add some occlusions to the facades to evaluate the performances of the proposed method. First, we collect and crop some occlusions from Cityscapes [4] in urban scenes. Then, we randomly choose an occlusion and paste it at a random location of a facade image. Finally, we use Poisson matting [37] to merge the occlusion and the facade image. Some occluded facade images can be seen in **Fig. 7**. We call this dataset ECP-occluded.

The ENPC Art-deco dataset, first used in [8], contains 79 rectified facade images cropped from the Art-deco style buildings in Paris. Similar to the ECP dataset, this dataset consists of seven classes: window, wall, balcony, door, roof, sky, and shop. The facades of the ENPC Art-deco dataset have some real occlusions that cause great challenges for facade parsing. We perform a 5-fold cross-validation on this dataset.

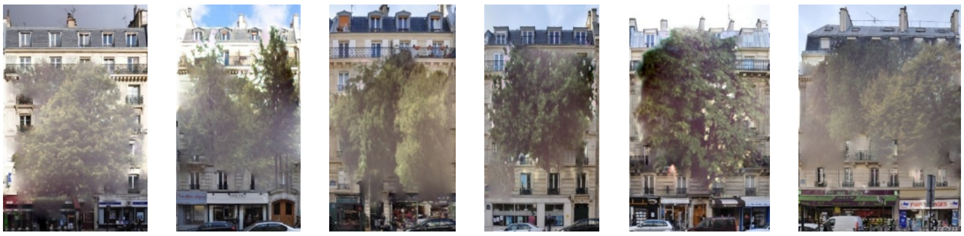


Fig. 7. The synthetic facade images on the ECP-occluded dataset.

Most facade images on the ENPC Art-deco dataset have clean facades or present sparse occlusions, such as scattered branches and leaves of trees, which are easy to parse. The synthetic images on the ECP-occluded dataset have large and dense occlusions which cause great challenges for facade parsing. The two datasets are complementary, and the synthetic ECP-occluded dataset can better reflect the anti-occlusion ability of methods.

Our experiments employ mostly used metrics in facade parsing to evaluate the performances, including the class average accuracy (class avg.), total pixel accuracy (total acc.) and mean intersection-over-union (mean IoU).

4.2 Implementation Details

We implement the proposed architecture based on TensorFlow and train it on a single GPU with a mini-batch size of 2. The whole training process only takes several hours. We adopt the ResNet50 [32] pre-trained on ImageNet as the backbone and fine-tune the weights with facade data. Other parameters are initialized by Xavier. The Adam optimizer with a basic learning rate of 1e-4 and weight decay of 0.0001 is used to optimize the whole network. To make the model robust, we adopt some data augmentation techniques: random scaling (from 0.5 to 2.0), random cropping (512 × 512) and random horizontal flipping. The source code will be published at https://github.com/wohaiyo/RPCNet in the near future.

4.3 Comparison with the State-of-the-art

4.3.1 On ECP-occluded Dataset

Experimental results on the ECP-occluded dataset are shown in **Table 1**. We mainly report the quantitative results of deep learning methods considering that these traditional methods perform poorly. Here, we compare with U-Net [18], which is popular in medical image segmentation and state-of-the-art semantic segmentation methods, such as ResNet50-FCN [32], PSPNet [20], Deeplabv3+ [21], and Pyramid ALKNet [26]. Our model consistently has better performance than these methods in all metrics. ResNet50-FCN [32] is a plain convolutional neural network architecture for semantic segmentation. It can only be able to perform pixel-wise classification in visible facade regions, which is difficult to deal with occlusions. Different from ResNet50-FCN, Deeplabv3+ [21] uses the atrous spatial pyramid pooling to obtain the context information, and Pyramid ALKNet [26] collects the long-range dependencies of deep features from the horizontal and vertical directions. DeepLabv3+

achieves better results than Pyramid ALKNet since it captures detailed visible information using low-level features. Both of them achieve better performances than ResNet50-FCN, but they also cannot predict the true labels in occluded regions of facades correctly. In contrast, our method solves these problems and achieves better performance. By capturing the nonlocal context information and considering the repetitive patterns of facades, our method can generate the discriminative features and infer contents in occluded regions by RPCNet which learns the characteristics from man-made structures.

Table 1. Quantitative comparisons (%) with state-of-the-art semantic segmentation methods on the ECP-occluded dataset.

Method	Class avg.	Total acc.	mean IoU
UNet [18]	62.2	75.5	51.5
ResNet50-FCN [32]	76.3	83.0	67.3
PSPNet [20]	75.9	82.8	66.7
DeepLabv3+ [21]	81.9	87.5	73.2
Pyramid ALKNet [26]	80.7	86.2	71.8
Ours	**82.6**	**87.7**	**73.7**

4.3.2 On ENPC Art-deco Dataset

On the ENPC Art-deco dataset, some facade images present real occlusions. We compare our method with previous state-of-the-art facade parsing methods and semantic segmentation methods. For fairness, the results of the facade parsing methods are from their original papers. Those networks for semantic segmentation are not trained and tested on this dataset. Here, we train them with the ENPC Art-deco dataset.

Table 2. Quantitative comparisons (%) with state-of-the-art methods on the ENPC Art-deco dataset using the splits of ours.

Method	Class avg.	Total acc.	mean IoU
Gadde et al. [8]	72.9	78.8	59.4
Cohen et al. [11]	78.1	85.3	-
Kozinski et al. [1]	83.7	88.8	-
Cohen et al. [12]	84.0	88.3	-
Autocontext [14]	84.8	89.0	73.5
UNet [18]	77.8	85.7	68.8
ResNet50-FCN [32]	86.2	89.4	79.2
PSPNet [20]	84.9	88.9	77.8
DeepLabv3+ [21]	89.3	92.0	83.2
Pyramid ALKNet [26]	89.3	91.9	83.1
Ours	**89.9**	**92.3**	**83.8**

In this real dataset, we first compare with some traditional methods for facade parsing, and then compare with some segmentation methods mentioned in **Table 1**. Compared with the traditional method [1] which employs an MRF shape prior to parse occluded facades, we adopt an RPCNet to infer the occluded regions by referring to the repetitive patterns in visible regions. The grammar requires careful definition by users and heavy computational costs at the parsing stage. However, our RPCNet can learn the potential rules through training without complex definitions, and our model runs faster and achieves better accuracies (see **Table 2**). [12] used the symmetries and repetitions of facades, but the hard constraints, such as strict symmetry, may hurt the accuracies of parsing results. Because the constraint of symmetry is not always met for many buildings. In contrast, our method decomposes the facade image into invisible and visible parts, and uses the visible repetitive patterns to infer the occluded contents which generally meets the structures of buildings and is also consistent with human reasoning. The experimental results in **Table 2** demonstrate that our repetitive pattern completion is reasonable and effective. Furthermore, due to discriminative features generated by the proposed context aggregation module and occluded contents inferred by the RPCNet, our method outperforms state-of-the-art semantic segmentation methods in all metrics.

Fig. 8 shows some visual comparisons with state-of-the-art methods. The accuracies of traditional methods are significantly lower than ours. Here, we present the visual results of the four deep learning-based approaches listed in **Table 2**. Different from the ECP-occluded dataset, the ENPC Art-deco dataset has real occlusions in front of building facades. The parsing results of ResNet50-FCN in visible facade regions look well, but terrible in occluded regions. This is because ResNet50-FCN cannot obtain structural information about facade elements from the appearances of occlusions. DeepLabv3+ and Pyramid ALKNet perform better than ResNet50-FCN using context information, but they fail in some serious occlusions. Specifically, the regular repetitions of the facade are broken and the shapes of windows and balconies are not maintained as rectangular. We can see that our method greatly improves the visual quality compared with state-of-the-art segmentation methods. Specifically, our method can encode the shapes and repetitive patterns of manmade structures into networks and solve these problems. Note that all of the parsing results are output directly from our model, without any post-processing.

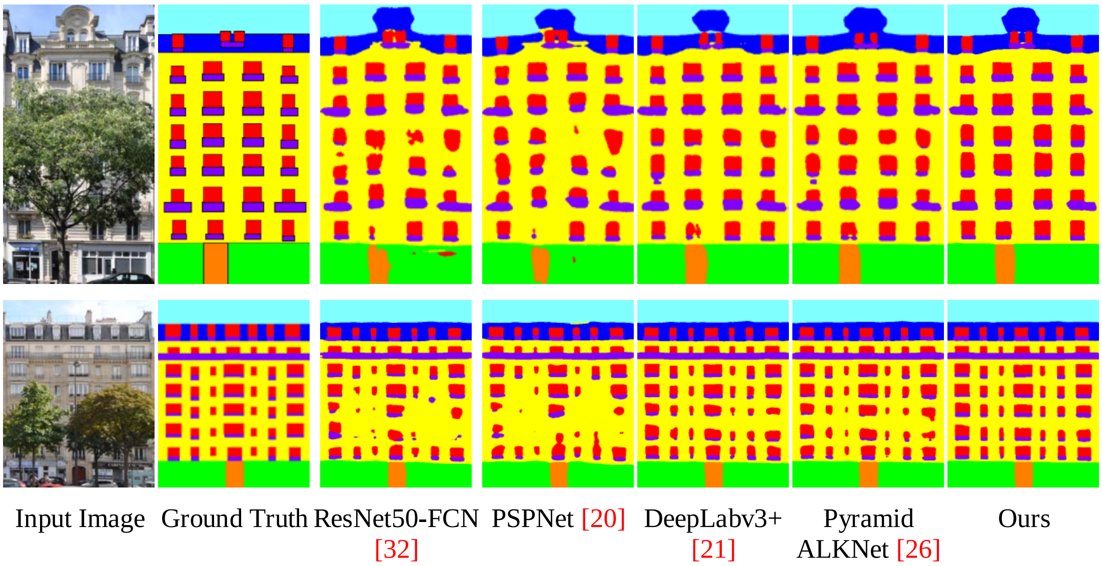


Fig. 8. Visual comparisons of facade parsing results on the ENPC Art-deco dataset.

DeepFacade [5] does not provide the source code and facade parsing results. For comparing with them, we test our method with their data splits on the ENPC Art-deco dataset. In **Table 3**, the experimental results of DeepFacade are from their paper. The class average accuracy is not objective enough to evaluate the segmentation results of the facade images. Some labels, such as doors and walls, are extremely unbalanced on the dataset. Therefore, the total accuracy is a better choice to compare the performances of different methods. Even building facade images with occlusions are rarely seen in their test data, our method achieves improvements in total accuracy and mean IoU, which illustrates that our model also has an advantage in parsing visible building facades.

Table 3. Quantitative comparisons (%) with state-of-the-art methods on the ENPC Art-deco dataset using the dataset splits of DeepFacade [5].

Method	**Class avg.**	**Total acc.**	**mean IoU**
DeepFacade [5]	**92.0**	92.9	79.8
Ours	90.3	**93.4**	**84.9**

4.4 Ablation Study

To further analyze the effect of the proposed context aggregation module and the repetitive pattern completion, we carry out ablation experiments on the ECP-occluded dataset and the ENPC Art-deco dataset. The baseline models are ResNet50-FCN and VGG16. As shown in **Table 4**, compared with the baseline model ResNet50-FCN, we adopt the repetitive pattern completion to determine the main occlusions and infer the contents using the regularities of repetitive patterns in visible parts. With only repetitive pattern completion, our model achieves a 2.2% improvement of mean IoU on the ECP-occluded dataset and a 1.2% improvement of mean IoU on the ENPC Art-deco dataset. Our context aggregation module can capture the structure information from multiple receptive fields. Our model with only context aggregation achieves better performances, a 6.1% improvement of mean IoU on the ECP-occluded dataset and a 4.4% improvement of mean IoU on the ENPC Art-deco dataset. Finally, with context aggregation and repetitive pattern completion, all metrics on the two datasets have improvements. These improvements mainly come from the layout regularization in occluded regions. To further show the benefits of the proposed key modules, we use VGG16 as a baseline model. Experiments in **Table 4** demonstrate that the context aggregation module and the repetitive pattern completion both have significant improvements in all metrics.

Table 4. Ablation study (%) of the proposed method. The results of the first and second rows are from the ECP-occluded dataset and the ENPC Art-deco dataset with ResNet50-FCN baseline model. The third and fourth rows are from the ECP-occluded dataset and the ENPC Art-deco dataset with the VGG16 baseline mode. “RP” means the repetitive pattern completion and “CA” means the proposed context aggregation module.

Method	**Class avg.**	**Total acc.**	**mean IoU**
Baseline (ResNet50-FCN)	76.3	83.0	67.3
+RP	79.0	84.7	69.5
+CA	82.1	87.5	73.4
+RP +CA	**82.6**	**87.7**	**73.7**
Baseline (ResNet50-FCN)	86.2	89.4	79.2
+RP	87.5	90.5	80.4
+CA	89.6	92.2	83.6
+RP +CA	**89.9**	**92.3**	**83.8**
Baseline (VGG)	74.2	80.5	64.7
+RP	77.6	83.2	67.5
+CA	81.5	86.8	72.5
+RP +CA	**82.7**	**87.3**	**73.3**
Baseline (VGG)	84.9	88.1	76.1
+RP	87.0	89.6	78.4
+CA	88.9	91.6	82.1
+RP +CA	**89.3**	**91.6**	**82.2**

Table 5. Ablation study (%) of the proposed method in occluded regions. We mainly report the IoU of the window, balcony and wall labels due to these labels occupy the most occluded regions. The results of the first and second rows are from the ECP-occluded dataset and the ENPC Art-deco dataset with ResNet50-FCN baseline model. The results of the third and fourth rows are from the ECP-occluded dataset and the ENPC Art-deco dataset with the VGG16 baseline model. "RP" means the repetitive pattern completion and "CA" means the proposed context aggregation module.

Method	Window	Wall	Balcony	mean IoU
Baseline (ResNet50-FCN)	38.3	74.2	45.7	52.8
+RP	52.9	77.2	52.2	60.8
+CA	56.7	82.5	62.5	67.2
+RP +CA	**59.7**	**83.0**	**62.6**	**68.4**
Baseline (ResNet50-FCN)	53.3	78.2	44.7	58.8
+RP	66.2	82.0	55.4	67.9
+CA	70.5	85.9	62.4	72.9
+RP +CA	**72.7**	**86.3**	**63.5**	**74.2**
Baseline (VGG)	27.2	69.6	32.6	43.1
+RP	48.8	74.8	43.8	55.8
+CA	53.9	81.5	57.7	64.4
+RP +CA	**58.2**	**82.5**	**61.0**	**67.2**
Baseline (VGG)	47.1	75.1	39.8	54.0
+RP	63.2	79.9	54.0	65.7
+CA	67.3	83.9	57.9	69.7
+RP +CA	**69.2**	**84.2**	**61.1**	**71.5**

Because the window, balcony and wall labels occupy most of the occluded regions, we also report the results of these three classes. In Table 5, we can see that all the metrics are improved on both datasets significantly with our specifically designed context aggregation module. Furthermore, our final model achieves the best performance on these three labels and mean IoU with a context aggregation module and repetitive pattern.

With the context aggregation module, our model has achieved significant improvements, but the layouts of occluded regions still have problems, resulting in lower accuracies of window and balcony labels. To generate regular element structures and orderly layouts of facades, which are very essential for facade parsing and its downstream applications, such as image inpainting and 3D semantic modeling. We use the repetitive pattern completion branch to refine the coarse facade parsing results. As shown in Tabel 5, with only the repetitive pattern completion, the metrics of window label are improved significantly in the first row. More importantly, the visual results are much better than those of methods simply learning to fit the data, which makes our method have much higher practical value.

Table 6. Quantitative comparisons with state-of-the-art methods on the ECP-occluded dataset under different occluded ratios.

Method	20~30%	30~40%	40~50%	50~60%
ResNet50-FCN [32]	78.2	74.7	71.2	67.3
PSPNet [20]	77.2	73.6	70.4	66.7
DeepLabv3+ [21]	80.4	77.9	75.6	73.2
Pyramid ALKNet [26]	80.7	77.7	75.5	71.8
Ours	**80.9**	**78.2**	**76.3**	**73.7**

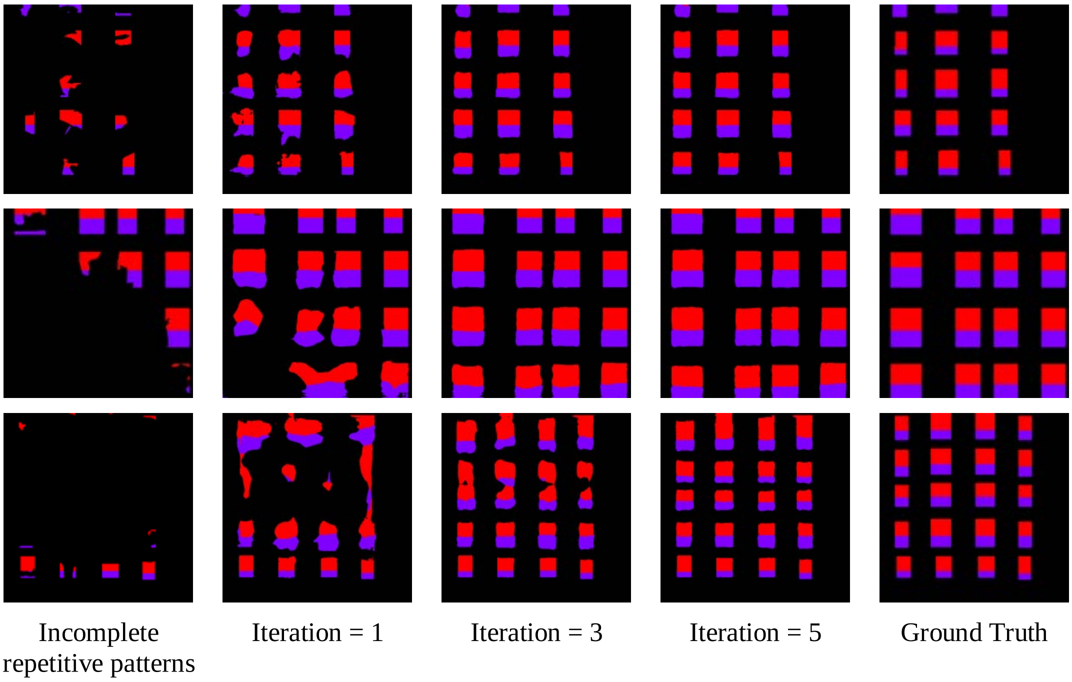


Fig. 9. Visual completion process of the repetitive pattern using RPCNet.

To further demonstrate the effectiveness of the proposed RPCNet, we evaluate the performance of the deep learning-based methods on the ECP-occluded dataset of different occluded ratios. In **Table 6**, with the occluded ratio increasing, the performance of some state-of-the-art methods, like DeepLabv3+ and Pyramid ALKNet, reduces significantly. Our method outperforms the others under different occluded ratios, especially under large ratios. We visualize some challenging completed repetitive patterns in **Fig. 9**. In the first row, the visible parts are not regular due to the arbitrary occlusions, and our RPCNet can reason for the layouts of facades and refine the shape of elements. In addition, the third row shows the more challenging case in which the input image is almost completely occluded. We can see that the RPCNet has difficulty generating reasonable structures even with three iterations. While using more iterations, the facade structures become better through the visible and reliable parts.

As it can be seen from **Table 7**, the proposed deep model with context aggregation module spends about 61ms to process a 512 × 512 image. The time cost increases by 5ms per iteration of RPCNet. Moreover, we also show the cross-entropy loss curves in the three different training stages. As shown in **Fig. 10**, the losses of training the RPCNet using synthetic repetitive patterns are higher than the others significantly. We apply the pre-trained model on real repetitive patterns of RPCNet and real facade datasets of multi-task learning. The proposed method converges quickly.

Table 7. The time cost per image of RPCNet with different iterations. “CA” means the proposed context aggregation module.

Method	**CA**	**iteration=1**	**iteration=3**	**iteration=5**
Time[ms]	61	67	76	86

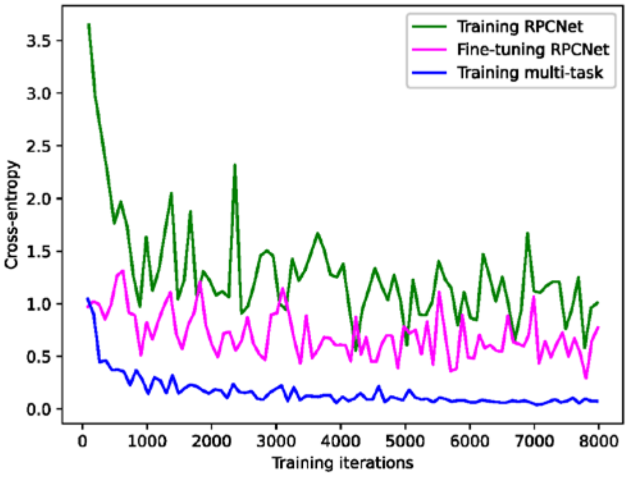


Fig. 10. The loss curves during training RPCNet, fine-tuning RPCNet and training multi-tasks.

4.5 Applications

Facade parsing results have many potential applications. Here, we apply our method in image inpainting and 3D semantic modeling and obtain promising results.

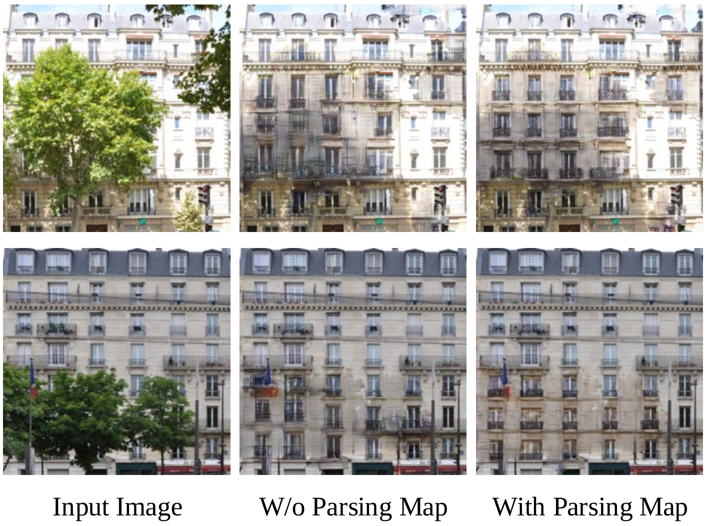


Fig. 11. Visual comparisons of inpainting results without/with the guidance of facade parsing results.

4.5.1 Image Inpainting

Occlusions heavily affect the appearances and structures of facades. It is inevitable to remove occlusions from facade images. Since we cannot obtain the ground truth of occluded regions on the ENPC Art-deco dataset, we use the clean ECP dataset to train the inpainting model and test on real images of the ENPC Art-deco dataset. Different from [38], we use the parsing results to guide the facade image inpainting process. As shown in **Fig. 11**, compared with the inpainting results without the guidance of parsing results, our results have a regular layout in occluded regions of the facade. For example, the shape and texture of the windows are more realistic. The visual inpainting comparisons further demonstrate that our method achieves good facade parsing results.

4.5.2 3D Semantic Modeling

The parsing results of the facade image can be straightforwardly used to build procedural modeling. As shown in **Fig. 12**, a facade image is fed into the proposed architecture to generate the corresponding parsing result. We encode the parsing result following a set of CGA rules in CityEngine [39]. In detail, we split the procedural modeling into layers in the vertical direction from bottom to top. In each layer, the facade is further subdivided into rows and columns that encoded by CGA rules. The solid and wireframe 3D semantic reconstruction results are shown in **Fig. 12**.

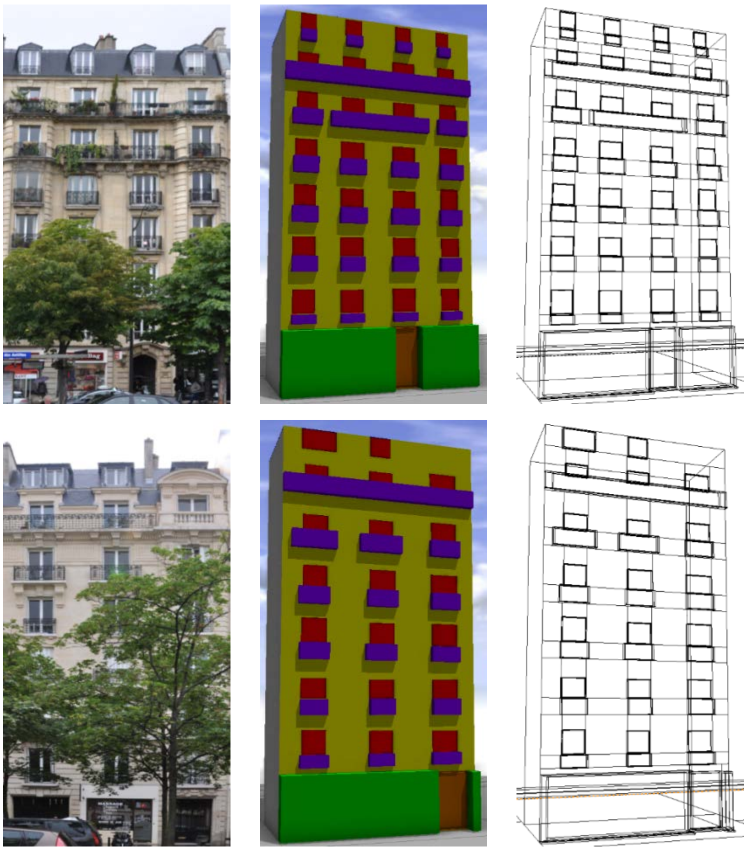


Fig. 12. 3D semantic modeling with our parsing result.

5. Conclusion

In this paper, for the first time, we presented an end-to-end deep network for facade parsing with serious occlusions. After decomposing a facade image into visible and invisible parts by using occlusion reasoning, the network adopts the context aggregation module to capture nonlocal information for the coarse semantic segmentation of visible parts. With the proposed repetitive pattern completion branch, the contents in occluded regions are well inferred by referring to the regularity in the visible part. The final parsing results are merged by a fusion module using visible and invisible parts. The overall end-to-end model is trained via multi-task learning.

Comprehensive experiments and ablation studies on real and synthetic datasets demonstrate the outstanding performance of the proposed method and its key modules. Moreover, we further applied the facade parsing results to facade image inpainting and 3D semantic modeling of buildings to verify their practical values.

Our method also has its limitations. Since the RPCNet relies too much on repetitive patterns, it tends to fail in some strange and extreme cases, such as buildings with irregular elements and layouts. In the future, we plan to handle these challenging scenes by investigating domain knowledge of building structure and more efficient network structure.

References

[1] M. Kozinski, R. Gadde, S. Zagoruyko, G. Obozinski, and R. Marlet, “A MRF shape prior for facade parsing with occlusions,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2820–2828, Jun. 2015. Article (CrossRef Link)

[2] N. Silberman, D. Hoiem, P . Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *Proc. of the European Conference on Computer Vision*, pp. 746–760, Oct. 2012. Article (CrossRef Link)

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231– 1237, Sep. 2013. Article (CrossRef Link)

[4] 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. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, Jun. 2016. Article (CrossRef Link)

[5] H. Liu, Y . Xu, J. Zhang, J. Zhu, Y . Li, and C. S. Hoi, “DeepFacade: A deep learning approach to facade parsing with symmetric loss,” *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3153-3165, Dec. 2020. Article (CrossRef Link)

[6] H. Riemenschneider, U. Krispel, W. Thaller, M. Donoser, S. Havemann, D. Fellner, and H. Bischof, “Irregular lattices for complex shape grammar facade parsing,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1640–1647, Jun. 2012. Article (CrossRef Link)

[7] J. Liu, E. Z. Psarakis, Y . Feng, and I. Stamos, “A kronecker product model for repeated pattern detection on 2d urban images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2266–2272, Jul. 2018. Article (CrossRef Link)

[8] R. Gadde, R. Marlet, and N. Paragios, “Learning grammars for architecture-specific facade parsing,” *International Journal of Computer Vision*, vol. 117, no. 3, pp. 290–316, Mar. 2016. Article (CrossRef Link)

[9] O. Teboul, “Ecole centrale paris facades database,” 2010. [Online]. Available: http://vision.mas.ecp.fr/Personnel/teboul/data.php

[10] O. Teboul, L. Simon, P . Koutsourakis, and N. Paragios, “Segmentation of building facades using procedural shape priors,” in *Proc. of the IEEE Conference on Computer Vision and Pattern*

Recognition, pp. 3105–3112, Jun. 2010. Article (CrossRef Link)

[11] A. Cohen, A. G. Schwing, and M. Pollefeys, “Efficient structured parsing of facades using dynamic programming,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3206–3213, Jun. 2014. Article (CrossRef Link)

[12] A. Cohen, M. R. Oswald, Y . Liu, and M. Pollefeys, “Symmetry-aware facade parsing with occlusions,” in *Proc. of the International Conference on 3D Vision*, pp. 393–401, Oct. 2017. Article (CrossRef Link)

[13] M. Mathias, A. Martinovi´ c, and L. V an Gool, “Atlas: A three-layered approach to facade parsing,” *International Journal of Computer Vision*, vol. 118, no. 1, pp. 22–48, May 2016. Article (CrossRef Link)

[14] R. Gadde, V . Jampani, R. Marlet, and P . V . Gehler, “Efficient 2d and 3d facade segmentation using auto-context,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1273–1280, Apr. 2017. Article (CrossRef Link)

[15] A. Wendel, M. Donoser, and H. Bischof, “Unsupervised facade segmentation using repetitive patterns,” in *Proc. of the Joint Pattern Recognition Symposium*, pp. 51–60, Sep. 2010. Article (CrossRef Link)

[16] C. Rodriguez-Pardo, S. Suja, D. Pascual, J. Lopez-Moreno, and E. Garces, “Automatic extraction and synthesis of regular repeatable patterns,” *Computers & Graphics*, vol. 83, pp. 33–41, Oct. 2019. Article (CrossRef Link)

[17] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 2017. Article (CrossRef Link)

[18] O. Ronneberger, P . Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 234–241, Oct. 2015. Article (CrossRef Link)

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017. Article (CrossRef Link)

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, Jun. 2017. Article (CrossRef Link)

[21] L.-C. Chen, Y . Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. of the European Conference on Computer Vision*, pp. 833-851, Sep. 2018. Article (CrossRef Link)

[22] M. Schmitz and H. Mayer, “A convolutional network for semantic facade segmentation and interpretation,” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLI-B3, pp.709-715, Jun. 2016. Article (CrossRef Link)

[23] R. Fathalla and G. Vogiatzis, “A deep learning pipeline for semantic facade segmentation,” in *Proc. of the British Machine Vision Conference*, pp. 120.1–120.13, Sep. 2017. Article (CrossRef Link)

[24] J. Femiani, W. R. Para, N. Mitra, and P. Wonka, “Facade segmentation in the wild,” *arXiv preprint arXiv:1805.08634*, 2018. Article (CrossRef Link)

[25] H. Liu, J. Zhang, J. Zhu, and S. C. Hoi, “DeepFacade: A deep learning approach to facade parsing,” in *Proc. of the 26th International Joint Conference on Artificial Intelligence*, pp. 2301–2307, Aug. 2017. Article (CrossRef Link)

[26] W. Ma, W. Ma, S. Xu, and H. Zha, “Pyramid ALKNet for semantic parsing of building facade image,” *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 6, pp. 1009-1013, June 2021. Article (CrossRef Link)

[27] W. Ma, S. Xu, W. Ma, and H. Zha, “Multiview feature aggregation for facade parsing,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2020. Article (CrossRef Link)

[28] W. Ma and W. Ma, “Deep window detection in street scenes,” *KSII Transactions on Internet and Information Systems*, vol. 14, no. 2, pp. 855–870, Feb. 2020. Article (CrossRef Link)

[29] C.-K. Li, H.-X. Zhang, J.-X. Liu, Y .-Q. Zhang, S.-C. Zou, and Y .-T. Fang, “Window detection in facades using heatmap fusion,” *Journal of Computer Science and Technology*, vol. 35, no. 4, pp.

900–912, Jul. 2020. Article (CrossRef Link)

[30] K. Bacharidis, F. Sarri, and L. Ragia, "3D building facade reconstruction using deep learning," *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 322, May 2020. Article (CrossRef Link)

[31] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. Article (CrossRef Link)

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Jun. 2016. Article (CrossRef Link)

[33] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, Jul. 2017. Article (CrossRef Link)

[34] P . Isola, J.-Y . Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, Jun. 2017. Article (CrossRef Link)

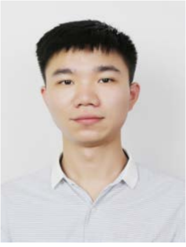
[35] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015. Article (CrossRef Link)

[36] C. Peng, X. Zhang, G. Y u, G. Luo, and J. Sun, "Large kernel matters– improve semantic segmentation by global convolutional network," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4353– 4361, Jun. 2017. Article (CrossRef Link)

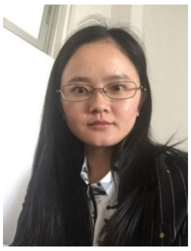
[37] P . Rez, M. Gangnet, and A. Blake, "Poisson image editing," *Acm Transactions on Graphics*, vol. 22, no. 3, pp. 313-318, p. 313, Jul. 2003. Article (CrossRef Link)

[38] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edge-connect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019. Article (CrossRef Link)

[39] ESRI, "Cityengine," 2016. [Online]. Available: http://www.esri.com/software/cityengine



Wenguang Ma was born in 1996. He received the B.S. degree in Computer Science and Technology from Beijing University of Technology, Beijing, China in 2018. He is currently pursuing the M.S. degree in Beijing University of Technology. His research interests include Computer Vision and Machine Learning.



Wei Ma received her Ph.D. degree in Computer Science from Peking University, in 2009. She is currently an Associate Professor at Faculty of Information Technology, Beijing University of Technology. Her current research interests include image/video repairing, image/video semantic understanding and 3D vision.

Shibiao Xu received the B.S. degrees in Information Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and the Ph.D. degree in Computer Science from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an associate professor in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include image based three-dimensional scene reconstruction and scene semantic understanding.