# DeepFacade: A Deep Learning Approach to Facade Parsing

**Hantang Liu[1], Jialiang Zhang[1], Jianke Zhu[1,2]*, Steven C.H. Hoi[3]**
[1] College of Computer Science, Zhejiang University, China
[2]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies
[3]School of Information Systems, Singapore Management University, Singapore
{liuhantang, zjialiang, jkzhu}@zju.edu.cn, chhoi@smu.edu.sg

## Abstract

The parsing of building facades is a key component to the problem of 3D street scenes reconstruction, which is long desired in computer vision. In this paper, we propose a deep learning based method for segmenting a facade into semantic categories. Man-made structures often present the characteristic of symmetry. Based on this observation, we propose a symmetric regularizer for training the neural network. Our proposed method can make use of both the power of deep neural networks and the structure of man-made architectures. We also propose a method to refine the segmentation results using bounding boxes generated by the Region Proposal Network. We test our method by training a FCN-8s network with the novel loss function. Experimental results show that our method has outperformed previous state-of-the-art methods significantly on both the ECP dataset and the eTRIMS dataset. As far as we know, we are the first to employ end-to-end deep convolutional neural network on full image scale in the task of building facades parsing.

## 1 Introduction

Building facades parsing is an important problem in computer vision. It enjoys many real world applications. First, this problem is key to the 3D reconstruction of street scenes, which has long been desired in computer vision community. Successful parsing of building facades can not only store building information more effectively but also record the information based on rules. These rules can be further used to reconstruct different styles of building facades, which is useful in game engines. Second, precise parsing of building facade can be useful in street map reconstruction and automatic driving cars. It can help the car to understand the surrounding environments better and increase security. The task's goal is to semantically identify each pixel's category. The semantic categories mainly consist building facades, like window, wall, balcony and so on.

*Corresponding author.

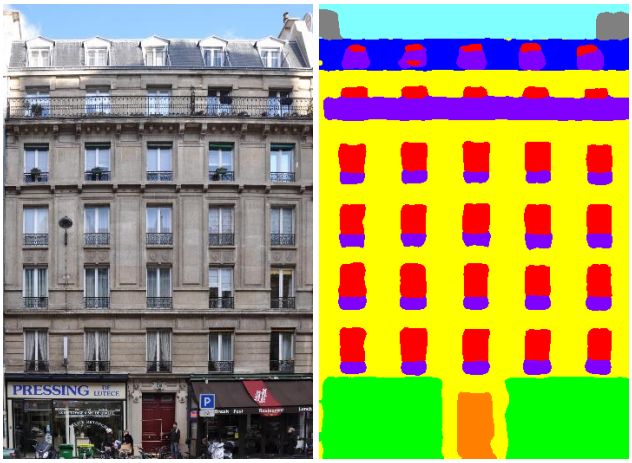



Figure 1: An example of the parsing result of our approach.

However, building facades parsing is a very challenging problem. This problem is usually formulated as an image segmentation problem. This problem is difficult not only due to the vast variation across different environments but also because of the changing in illumination, visual perspective, and occlusions. First, buildings are usually in a wild complex environment. Building styles also vary significantly across different areas. Even in the same city, there are no two identical buildings. The forming components of building facade also have great diversity, like texture, genre, and color. Second, there is also a lot of occlusions. Third, the parsing of building facades may involve some complex non-building elements, like cars and pedestrians. The left image in Figure 1 shows this difficulty.

In literature, building facades parsing has been actively studied and various methods have been proposed in computer vision [Mathias *et al.*, 2016; Cohen *et al.*, 2014]. Most of these methods operate on per-pixel or super-pixel level, addressing this problem as an image segmentation problem. Early methods [Teboul *et al.*, 2010] assume that building facades have an appropriate shape grammar. This poses strong prior knowledge on the facade of a building. If the prior does not apply, the methods fail. More recent methods [Mathias *et al.*, 2016] tried to learn label information at different abstract levels. Despite achieving some promising results,

the learning process is still hand crafted to some extent and may not perform very well in some scenarios. [Cohen *et al.*, 2014] took another approach and used dynamic programming. The optimization technique increased the performance significantly, however, it does not make full use of learning technique to learn from the data.

Recently, deep learning has shown its power in various computer vision tasks, like image classification [Krizhevsky *et al.*, 2012], image recognition, object detection [Girshick *et al.*, 2014] and image segmentation [Long *et al.*, 2015; Chen *et al.*, 2016]. Even low-level image processing problems have also benefitted from deep learning techniques, like image denoising [Xie *et al.*, 2012], art style transfer [Gatys *et al.*, 2015] and image super-resolution [Dong *et al.*, 2014]. Deep learning has outperformed traditional vision approaches in a lot of benchmarks. As to the problem of image segmentation, deep learning has mainly seen its application in general image segmentation problems. [Schmitz and Mayer, 2016] has applied deep learning to facade parsing by treating the facade parsing as a general image segmentation problem. Despite promising results, the general technique for image segmentation has not delved into the specific problem of facade parsing to fully take advantage of the characteristics of this problem. Specifically, the man-made rules of the structures are not incorporated into the network.

While many previous methods have relied on hand crafted priors to parse the building facades, we explore a deep learning based approach to resolving the facade parsing problem. We call our method DeepFacade. As far as we now, we are the first to apply deep learning to facade parsing on full image scale. In particular, we present a novel symmetric regularizer to train the neural network to make use of both the learning capacity of deep convolutional neural networks and man-made rules in building facades.

The basic idea of our proposed symmetric neural facade parsing is to train deep convolutional neural networks with the constraints under man-made rules. The main focus of facade parsing problem is on {*window, door, balcony*}. These objects often have a high level of symmetry inside them. We impose a symmetric regularization on the aforementioned classes during training the network. As most windows have a square shape, we assume that bounding boxes generated by object detection will also be helpful to location and refine the shape of the predicted results. In particular, we use Region Proposal Network to generate the bounding boxes. We conducted experiments on ECP dataset and eTRIMS dataset. On the ECP dataset, our method outperforms the state-of-the-art by more than $6\%$ percent. On the eTRIMS dataset, our method outperforms the state-of-the-art by more than $10\%$ percent.

## 2 Related Work

The problem of building facade parsing has long been actively studied and there exists a lot of work on how to tackle this problem. [Zhao *et al.*, 2010] proposed an approach that parses registered images captured at ground level into architectural units for large-scale city modeling. Each parsed unit has a regularized shape, which can be used for further modeling purposes. [Wendel *et al.*, 2010; Recky *et al.*, 2011] used repetitive patterns to tackle the same problem. [Mathias *et al.*, 2011] took another approach by proposing an algorithm which automates this process through classification of architectural styles from facade images. Their classifier first identifies the images containing buildings, then separates individual facades within an image and determines the building style.

Many approaches assume a procedural grammar. [Müller *et al.*, 2007] combine the procedural modeling pipeline of shape grammars with image analysis to derive a meaningful hierarchical facade subdivision. [Ripperda and Brenner, 2006] use a process based on reversible jump Markov Chain Monte Carlo (rjMCMC) to guide the application of derivation steps during the construction of the tree. [Han and Zhu, 2009] study an effective top-down/bottom-up inference algorithm for parsing images. [Teboul *et al.*, 2011] address shape grammar parsing for facade segmentation using Reinforcement Learning (RL). Their methods achieve good results with a significant speed-up compared to previous methods. Shape priors may provide good regularization if the building facades are constructed under these grammars.

[Mathias *et al.*, 2016] propose a parsing method that consists of three distinct layers. In the first layer, facade labeling is learned at super-pixel level via Recursive Neural Network. In the middle layer, they introduce the knowledge about distinct facade elements. They combine the output of the RNN with object detectors. They model the merging procedure as a 2D Markov Random Field over the pixels. The MRF is solved via graph cut. The top layer encodes a set of rules and lead to a more structured configuration.

[Cohen *et al.*, 2014] present an optimization problem for which they can construct optimality certificates while being more efficient if not interested in their computation. They use a dynamic programming algorithm with extensions for improved expected case efficiency. The proposed algorithm requires individual executions of a dynamic program to find a labeling. Global optimality certificates are obtained if the individual algorithms remain independent.

Deep learning has shown its amazing power in various vision tasks. There has also been quite a body of work addressing the image segmentation problem. [Long *et al.*, 2015] is the first to train an end-to-end deep convolutional neural network for general image segmentation task. [Chen *et al.*, 2016] use dilated convolution instead of plain convolution. This approach avoids the use of a deconvolution layer, thus making the network easier to train. CRF post-processing can be applied to refine the results. However, the deep learning methods mainly focus on the network structure [Long *et al.*, 2015; Chen *et al.*, 2016] or the learning methodology, like batch normalization [Ioffe and Szegedy, 2015] or new initialization methods [He *et al.*, 2015]. Few works has looked into how to guide the neural network with prior information or assumptions. For the problem of facade parsing, [Schmitz and Mayer, 2016] has taken a fully convolutional network approach. They trained the network on facade image patches and did not take advantage of the structure of facades.

In this paper, we propose a novel symmetric loss for the deep convolutional neural network and demonstrate its efficacy on two facade parsing datasets.

# 3 Approach

To incorporate the man-made rules into the end-to-end system of a deep convolutional neural network, we propose a new loss term based on the common symmetry found in structures like windows, walls, and doors. Besides segmentation technique, we also found deep learning based detection helpful in the parsing procedure. This step is optional alongside the end-to-end segmentation pipeline.

## 3.1 Network Structure

Typically, a deep convolutional neural network consists of $l$ layers, each layer applies a linear convolution to its input, followed by an activation layer. Convolutional layers are usually followed by a pooling layer to downscale the feature map so that the final output has a smaller feature size for classification. A fully convolutional neural network replace the fully connect layers of a classification network with fully convolutional layers, making the network produce dense classification for each pixel in the last layer of the response map. We need to upsample the feature map to obtain an output that has the same size with the input image. Several methods can be used to achieve this goal, for example, bilinear filter and transposed convolution. Transposed convolution is called deconvolution in the FCN paper[Long *et al.*, 2015]

For the network structure, we follow the settings of FCN-8s in[Long *et al.*, 2015]. The first 13 layers of VGG16[Simonyan and Zisserman, 2014] are used as the base network. The two fully connected layers of the VGG16 network are cast into two fully convolutional layers. As the casted convolutional layer has a filter size of $1 \times 1$, the parameters of the fully connected layers in VGG16 can be directly copied to the fully convolutional layers. To unify the number of neurons, the two convolutional layers are set to have 4096 channels.

Transposed convolution (i.e., deconvolution in the FCN paper) can be used to upscale the response map, thus obtaining a prediction of the same size with the original image. FCN-32s directly upscale the feature 32 times to the original input size, so the final prediction may be coarse and will not be very accurate. Intermediate features in the early convolutional layers are also helpful in dense classification tasks, so we want to incorporate this information into the final segmentation phase. Also, upscaling the feature map gradually may result in a more accurate shape. FCN-8s first upscale the feature map to twice as large, then concatenate the upscaled feature map with the feature map after pool4. The new feature map is again upscaled to twice larger and is concatenated with the feature map after pool3. Then it is upscaled 8 times larger to the original image size.

## 3.2 A Symmetric Constraint for Man-Made Architectures

Typically, the segmentation network is trained with a regular cross entropy loss as follows

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

Here $\mathbf{x}$ is the input image array, $\mathbf{y}$ is the probability distribution of the category label of the image. $N$ is the number of pixels in the image, $i$ is pixel index. $y_i$ is the true probability distribution and $\hat{y}_i$ is the predicted probability distribution.

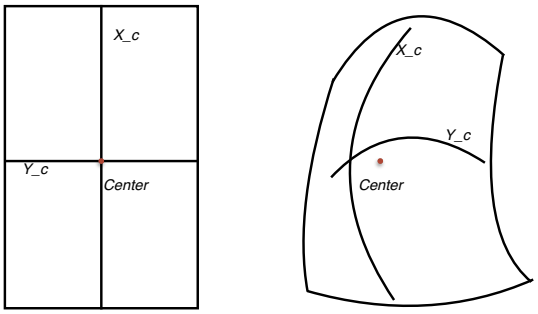



Figure 2: An illustration for the symmetric loss term. The centers of a symmetric object should fall on the center of both vertical and horizontal centers of the line segments.

Although deep convolutional neural networks have shown their strength in the problem of general image segmentation, this loss function does not pose any constraint or guide on the neural network. We expect the neural network to be able to utilize the man-made rules in building facades. So we propose the following symmetric loss term.

For each category that we want to impose the symmetric constraint, let $p$ denote each object that belongs to this category. As in the case of building facades, no two different objects of the same class have an intersection, it is feasible to partition the objects into a set $\mathbf{P}$ in which no two elements intersect. Then for each object $p$, the symmetric loss term is as follows:

$$\tilde{\mathcal{L}_s}(\mathbf{x}) = \sum_{p} (\mathrm{Var}[X_c] + \mathrm{Var}[Y_c]) \quad (2)$$

Let $\tilde{\mathcal{L}_s}(\mathbf{x})$ denote the overall symmetric loss for all the object $p \in \mathbf{P}$. Let $\mathcal{L}_s$ denote the symmetric loss for a single object $p$. Here $X_c$ is the random variable that represents the center of each horizontal line segment of object $p$, and $Y_c$ is the random variable that represents the center of each vertical line segment of object $p$.

Each object $p$ is a set of pixels $p = \{(x, y)\}$ where $x$ and $y$ are vertical and horizontal coordinates respectively. Let $x_{cj}$ be a sample of $X_c$, $x_{cj}$ represents the center of the $j$-th vertical line segment of $p$

$$x_{cj} = \frac{1}{N_j} \sum_{y=j}^{N_j} x \quad (3)$$

where $N_j$ is the number of pixels in vertical line segment $p_{vj} = \{(x, y) | y = j\}$. Similarly for horizontal symmetry we have

$$y_{ci} = \frac{1}{N_i} \sum_{x=i}^{N_i} y \quad (4)$$

Where $N_i$ is the number of pixels in horizontal line segment $p_{hi} = \{(x, y) | x = i\}$.

Ideally, for all $i$, $x_{cj}$ should have the same value because of vertical symmetry. The same goes for $y_{ci}$. However, inaccurate segmentation will lead to variance of the distribution

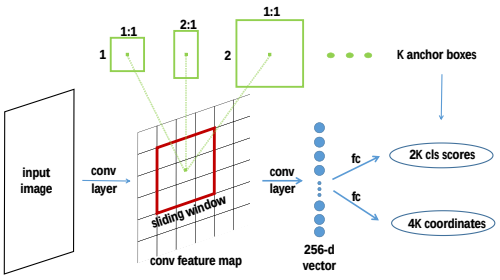



Figure 3: Bounding boxes example from Region Proposal Network (RPN)

Figure 4: Some examples of the region proposals generated by RPN.

of $X_c$ and $Y_c$, as illustrated in Figure 2. If an object $p$ is perfectly symmetric, the variance of $X_c$ and $Y_c$ should both be 0 as shown in the picture. Otherwise, the variance gets bigger as one direction goes further away from being symmetric. Let $\mathcal{L}_{sv}$ and $\mathcal{L}_{sh}$ denote the symmetric loss along the vertical and horizontal direction respectively. Then we have

$$\begin{aligned}\mathcal{L}_{sv} &= \frac{1}{N_j}\sum(x_{cj} - \frac{1}{N_j}\sum x_{cj})^2 \\ &= \mathrm{Var}[X_c]\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{sh} &= \frac{1}{N_i}\sum(y_{ci} - \frac{1}{N_i}\sum y_{ci})^2 \\ &= \mathrm{Var}[Y_c]\end{aligned} \tag{6}$$

We combine both terms to form a symmetry loss in both directions

$$\begin{aligned}\mathcal{L}_s &= \mathcal{L}_{sv} + \mathcal{L}_{sh} \\ &= \mathrm{Var}[X_c] + \mathrm{Var}[Y_c]\end{aligned} \tag{7}$$

The final loss function becomes

$$\tilde{\mathcal{L}} = \mathcal{L}(\mathbf{x}, \mathbf{y}) + \eta\tilde{\mathcal{L}}_s \tag{8}$$

where $0 \leq \eta \leq 1$. We call this loss function the symmetric loss in the following sessions.

### 3.3 Boosting the Performance Using Object Detection

Window, door, balcony are the most important structures in facade parsing. Most of these also have a square shape, we can use this prior information to greatly improve the visual results of our parsing. Object detection generates bounding boxes to show the location of a particular object. If the predicted bounding boxes can match the location of these objects well, then the results will be significantly refined. Here we use the Faster R-CNN [Ren *et al.*, 2015] to generate bounding boxes for windows.

The RPN was a region proposal generator in Faster R-CNN[Ren *et al.*, 2015], which is class-agnostic in multi-category. For single-category detection as ours (window), RPN is naturally a detector. The specification of RPN in our task is as follows.

Following [Ren *et al.*, 2015], a VGG-16 net pre-trained on the ImageNet dataset [Deng *et al.*, 2009] is adopted as base network. Our RPN is built on the top of $Conv5\_3$, then an intermediate $3 \times 3$ convolutional layer and two siblings $1 \times 1$ convolutional layers for classification and bounding box regression follows (more details in [Ren *et al.*, 2015]). In this way, feature stride of feature map 5 ($Conv5\_3$) is 16. We adopt 9 anchor boxes with 3 aspect ratios of $1:1, 1:2$ and $2:1$, and with box areas of $32^2, 64^2$ and $128^2$ which are different from the original RPN [Ren *et al.*, 2015] of $128^2$, $256^2, 512^2$ for the reason that the size of windows in the ECP dataset are generally small (ranging from $10^3$ to $10^4$ pixels) with respect to image size of MaxLength = 1000.

To choose a single bounding box in every possible position, we simply take two steps of NMS (Non-Maximum Suppression). First, NMS with threshold 0.7 is set to get Top-$M$ possible windows where $M$ is 100 in our experiments, then a threshold of 0.01 is set to suppress overlaps to take final proposals. 5-fold cross-validation is taken to generate proposals from all images. Some of the results are shown in Figure 4.

To apply the detection results to the segmentation, we first cast the bounding boxes into a score map across the whole image. Let $\mathbf{s} = \{s_{ijk}\}_{H\times W\times K}$ denote the score map for the prediction of an image. $\mathbf{s}$ is a matrix of size $H \times W \times K$, $H$ and $W$ are the height and width of the image, and $K$ is the number of semantic categories. Each entry $s_{ijk}$ represents the probability of pixel $(i, j)$ belonging to class $k$. In the case of facade parsing, the following classes are suitable for a detector to generate bounding boxes for them: {*window, door, balcony, chimney, shop*}. After generating bounding boxes for these classes in an image we have

$$\mathbf{s}_{ijk} = \begin{cases} 1, & \text{if } (i,j) \in \text{class-k} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Let $\mathbf{s}_{fcn}$ be the score map produced by FCN, then the final score will be a linear combination of the two

$$\hat{\mathbf{s}} = \mathbf{s}_{fcn} + \mathbf{s} \tag{10}$$

the predicted labels for each pixel $(i, j)$ will be

$$\hat{\mathbf{y}} = \mathrm{argmax}\,\hat{\mathbf{s}} \tag{11}$$

### 3.4 Training and Inference

First, we initialize the network with a model pre-trained on ImageNet data. We then apply the Adam optimizer to fine-tune the network. The initial learning rate is set to $10^{-6}$. For

the ECP dataset, training epoch is 100 and for the eTRIMS dataset training epoch is 80. The number is chosen by observing when the training loss stops decreasing. Dropout [Srivastava *et al.*, 2014] is used during training to prevent overfitting. The training procedure is the same for FCN-8s with a cross entropy loss and with a symmetric loss. The only difference is the loss function. For $\eta$ in Equation 8, we set it to be $0.17$ empirically.

Once the network is successfully trained, it can output a prediction directly given an input image. To increase the accuracy of some main classes in building facades and improve the visual quality, one can use the RPN bounding boxes to refine the segmentation result output by the network as described in Section 3.3

# 4 Experiments

## 4.1 Dataset

We evaluate our porposed approach on two different datasets, the Ecole Centrale Paris (ECP) Facades dataset [Teboul *et al.*, 2010] and the eTRIMS [Korč and Förstner, 2009] database.

**ECP Dataset**. The ECP dataset consists of 104 images of building facades. The dataset contains the following classes: {*window, wall, balcony, door, shop, sky, chimney, roof*}. All the images in the ECP dataset contains rectified and cropped facades of Haussmannian style buildings in Paris. The original annotation labeled the images using a Haussmannian-style grammer. This often results in imprecise or even wrong annotations. So we use the annotation provided by [Mathias *et al.*, 2016], where the annotation better fits the ground truth.

**eTRIMS Database**. The eTRIMS database has two variants. We use the 8-Class eTRIMS Dataset with 8 annotated object classes, which consists of 60 annotated images. The eTRIMS database consists of the following classes: {*Window, Wall, Door, Sky, Pavement, Vegetation, Car, Road*}. Different from the ECP dataset, images in the eTRIMS are not rectified. The windows and walls may not be a perfect square in most cases. This poses some challenge in the RPN refinement.

## 4.2 Experiment Setup

For the ECP dataset, we report three results based on the method configuration. The first is the FCN-8s trained with a regular softmax cross entropy loss, in the following sections it is denoted as $Ours^1$. The second is the FCN-8s trained with the symmetric loss proposed in Section 3.2, denoted as $Ours^2$. The third one is the results refined with RPN, denoted as $Ours^3$. We compare with dataset baseline and state-of-the-art results. in [Mathias *et al.*, 2016; Cohen *et al.*, 2014]. For the eTRIMS dataset, we report one result obtained by the FCN-8s trained with the symmetric loss. Objects in this dataset are often not rectangles, so in many cases bounding boxes cannot give a precise pixel-wise prediction. As a consequence, refinement by detection could not help too much and even may decrease the performance with some parameter settings. For the eTRIMS dataset, we also compare our results with state-of-the-arts methods [Yang and Förstner, 2011; Mathias *et al.*, 2016; Cohen *et al.*, 2014; Schmitz and Mayer, 2016].

| Class | [1] | [2] | [3] | [4] | Ours |
|---|---|---|---|---|---|
| Building [%] | 71 | 91 | 91 | 83 | **96.03** |
| Car [%] | 35 | 74 | 70 | - | **94.20** |
| Door [%] | 16 | 50 | 18 | **97** | 80.66 |
| Pavement [%] | 22 | 15 | 33 | - | **84.81** |
| Road [%] | 35 | 73 | 57 | - | **90.58** |
| Sky [%] | 78 | 97 | 97 | - | **98.06** |
| Vegetation [%] | 66 | 87 | 90 | - | **94.16** |
| Window [%] | 75 | 73 | 71 | 86 | **90.91** |
| total acc. [%] | 65.8 | 83.39 | 83.84 | 85 | **94.15** |

Table 1: Pixel accuracies on the eTRIMS dataset. Accuracies are shown in percentage. [1] is [Yang and Förstner, 2011], [2] is [Mathias *et al.*, 2016], [3] is [Cohen *et al.*, 2014], [4] is [Schmitz and Mayer, 2016]

## 4.3 Quantitative Evaluation

Table 2 shows the comparison result of our method and state-of-the-art methods on the ECP dataset. Ours trained with symmetric loss and then refined by RPN bounding boxes have beaten previous state-of-the-art methods by $5.06\%$ absolute percentage. We report three configurations of our method. $Ours^1$ is the result of FCN-8s trained with a plain cross entropy loss. $Ours^2$ is the result of FCN-8s trained with a symmetric loss. $Ours^3$ is the result of $Ours^2$ refined by the RPN bounding boxes. [1] is a baseline method. [2] is the result of layer 2 in [Mathias *et al.*, 2016] and [3] is the result of layer 3 in [Mathias *et al.*, 2016]. These two layers hold the top accuracies of this methods. [4], [5] and [6] are the three parameter setting reported in [Cohen *et al.*, 2014].

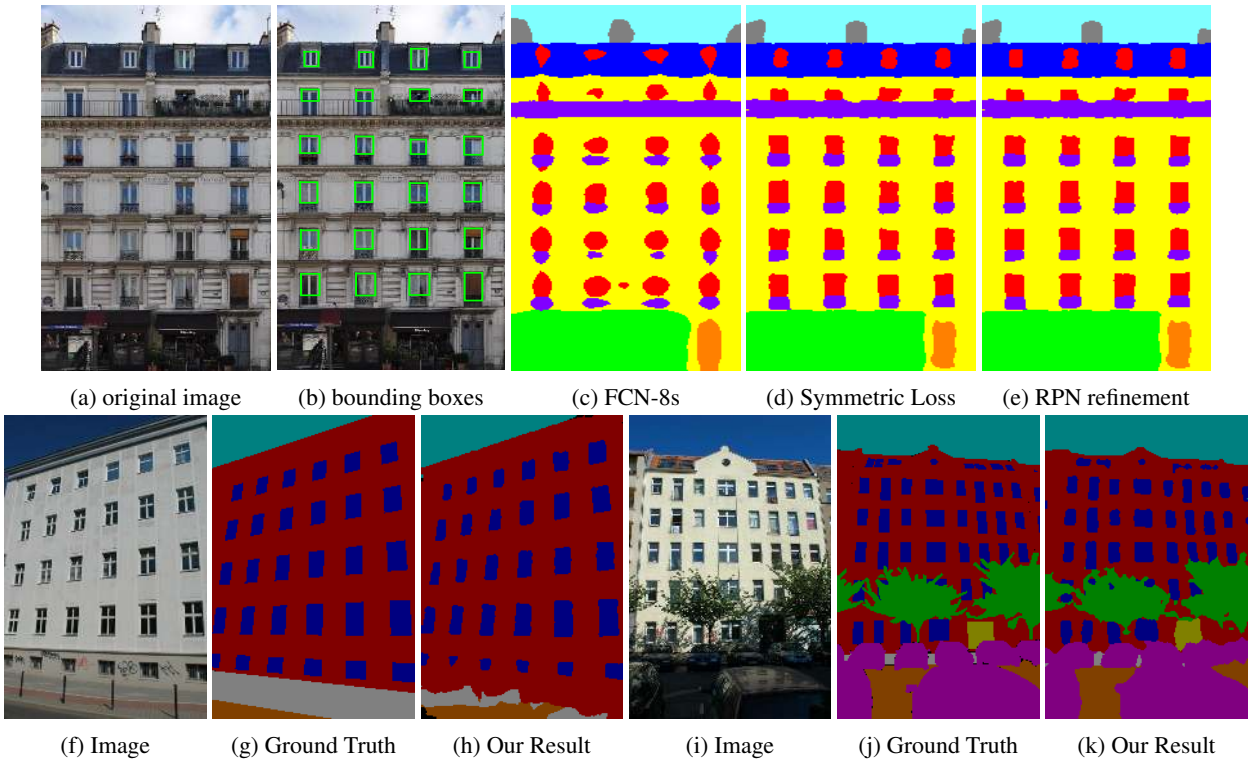
Our method not only outperforms previous state-of-the-art methods in total accuracy by a large percentage but also outperforms previous state-of-the-art methods in every single class. Specially, *window* is one of the most important classes in building facade parsing. Previous methods' best accuracy was $87\%$, our best result is $93.04\%$, that is more than $6\%$ percent improvement. As we can see in the table, {*door, balconey*} are hard classes compared to other easier classes. Our method also achieves a performance gain of $8.95\%$ and $3.07\%$ respectively.

As we can see, the accuracy of the *window* class increased from $86.81\%$ to $88.52\%$ after training with the symmetric loss proposed in this paper. This proves the efficacy of this regularization term. The accuracies of several classes get a performance boost one step further after being refined with RPN bounding boxes.

Table 1 shows the accuracies on the eTRIMS dataset. $Ours$ represents our method training the network using the symmetric loss. No bounding box refinement is applied to the output result. We can see again that our method beats the previous method in both total accuracy and every single class by a large extent. Especially in some hard cases like {*door, pavement*}. The overall performance gain is over $10\%$.

## 4.4 Qualitative Evaluation

Figure 5 shows four samples of qualitative result on the eTRIMS dataset. In each of the group, the left column shows the original image. The middle column shows the ground truth label. The right column shows the result obtained with our symmetric loss. Generally, the segmentation is visually pleasing. The location and shape of the objects in the im-



(a) original image (b) bounding boxes (c) FCN-8s (d) Symmetric Loss (e) RPN refinement

(f) Image (g) Ground Truth (h) Our Result (i) Image (j) Ground Truth (k) Our Result

Figure 5: Qualitative examples on the ECP and eTRIMS dataset.

| Class [%] | [1] | [2] | [3] | [4] | [5] | [6] | $Ours^1$ | $Ours^2$ | $Ours^3$ |
|---|---|---|---|---|---|---|---|---|---|
| Window [%] | 62 | 76 | 78 | 68 | 87 | 85 | 86.81 | 88.52 | **93.04** |
| Wall [%] | 82 | 90 | 89 | 92 | 88 | 90 | 96.08 | 95.79 | **96.14** |
| Balcony [%] | 58 | 81 | 87 | 82 | 92 | 91 | 92.44 | 94.64 | **95.07** |
| Door [%] | 47 | 58 | 71 | 42 | 82 | 79 | 86.08 | 85.17 | **90.95** |
| Roof [%] | 66 | 87 | 79 | 85 | 92 | 91 | 92.75 | **94.02** | 93.73 |
| Sky [%] | 95 | 94 | 96 | 93 | 93 | 94 | 96.62 | 97.48 | **97.72** |
| Shop [%] | 88 | 97 | 95 | 94 | 96 | 94 | **95.68** | 94.22 | 95.62 |
| Chimney [%] | - | - | - | 54 | 90 | 85 | 85.34 | **91.30** | 90.29 |
| total acc. [%] | 74.71 | 88.07 | 88.02 | 86.71 | 89.90 | 90.34 | 93.79 | 94.59 | **95.40** |

Table 2: Pixel accuracies comparison on the ECP dataset. [1] is [Yang and Förstner, 2011], [2] and [3] are two variants of [Mathias *et al.*, 2016], [4][5][6] are three variants of [Cohen *et al.*, 2014].

age are precisely predicted. Specially, windows get well predicted shapes and are generally symmetric.

Figure 5 shows two examples of qualitative comparison of different settings of our method. In each row, the left two images show the original building image and the detected bounding boxes on the windows. The third image shows the segmentation result of FCN-8s trained with a plain cross entropy loss. The fourth image shows the result trained with our symmetric loss. The image on the right shows the result refined with RPN bounding boxes. We can see the symmetric loss greatly improve the visual quality of the *window* class, making the output shape more symmetric, though not perfect. After refinement with RPN bounding boxes, the quality of the segmentation is further improved. The windows are more square and the edges of the windows are more smooth.

## 5 Conclusion

In this paper, we applied deep convolutional neural network to the 2D facade parsing problem. As far as we know, we are the first to train an end-to-end deep convolutional neural network to tackle this problem. We propose a symmetric regularization term and obtain a novel loss function for training the neural network. The symmetric regularization can help neural networks to predict the location and shapes of several object classes more precisely. We also propose an approach to refine the segmentation result using the Region Proposal Network (RPN). Experimental results on two challenging datasets have outperformed previous state-of-the-art methods significantly, proving the efficacy of our proposed approach.

## Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2016YFB1001501). This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

## References


[Chen *et al.*, 2016] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[Cohen *et al.*, 2014] Andrea Cohen, Alexander G. Schwing, and Marc Pollefeys. Efficient structured parsing of facades using dynamic programming. June 2014.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.

[Gatys *et al.*, 2015] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[Han and Zhu, 2009] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing with attribute grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):59–73, 2009.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[Korč and Förstner, 2009] F. Korč and W. Förstner. eTRIMS Image Database for interpreting images of man-made scenes. Technical Report TR-IGG-P-2009-01, April 2009.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[Mathias *et al.*, 2011] Markus Mathias, Andelo Martinovic, Julien Weissenberg, Simon Haegler, and Luc Van Gool. Automatic architectural style recognition. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3816:171–176, 2011.

[Mathias *et al.*, 2016] Markus Mathias, Andelo Martinovic, and Luc Van Gool. ATLAS: A three-layered approach to facade parsing. *International Journal of Computer Vision*, 118(1):22–48, 2016.

[Müller *et al.*, 2007] Pascal Müller, Gang Zeng, Peter Wonka, and Luc Van Gool. Image-based procedural modeling of facades. *ACM Transactions on Graphics (TOG)*, 26(3):85, 2007.

[Recky *et al.*, 2011] Michal Recky, Andreas Wendel, and Franz Leberl. Façade segmentation in a multi-view scenario. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 358–365. IEEE, 2011.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Ripperda and Brenner, 2006] Nora Ripperda and Claus Brenner. Reconstruction of façade structures using a formal grammar and rjmcmc. In *Joint Pattern Recognition Symposium*, pages 750–759. Springer, 2006.

[Schmitz and Mayer, 2016] Matthias Schmitz and Helmut Mayer. A convolutional network for semantic facade segmentation and interpretation. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 709–715, 2016.

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

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[Teboul *et al.*, 2010] Olivier Teboul, Loic Simon, Panagiotis Koutsourakis, and Nikos Paragios. Segmentation of building facades using procedural shape priors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3105–3112. IEEE, 2010.

[Teboul *et al.*, 2011] Olivier Teboul, Iasonas Kokkinos, Loic Simon, Panagiotis Koutsourakis, and Nikos Paragios. Shape grammar parsing via reinforcement learning. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2273–2280. IEEE, 2011.

[Wendel *et al.*, 2010] Andreas Wendel, Michael Donoser, and Horst Bischof. Unsupervised facade segmentation using repetitive patterns. In *Joint Pattern Recognition Symposium*, pages 51–60. Springer, 2010.

[Xie *et al.*, 2012] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012.

[Yang and Förstner, 2011] Michael Ying Yang and Wolfgang Förstner. Regionwise classification of building facade images. In *Photogrammetric image analysis*, pages 209–220. Springer, 2011.

[Zhao *et al.*, 2010] Peng Zhao, Tian Fang, Jianxiong Xiao, Honghui Zhang, Qinping Zhao, and Long Quan. Rectilinear parsing of architecture in urban environment. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 342–349. IEEE, 2010.