Magic-wall: Visualizing Room Decoration

Ting Liu
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044

Yunchao Wei*
National University of Singapore

Yao Zhao
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044

Si Liu
State Key Laboratory of Information Security, Institute of Information Engineering, CAS Beijing, China 100093

Shikui Wei†
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044
shkwei@bjtu.edu.cn

Figure 1: Given an image of an indoor scene, the proposed Magic-wall is able to automatically replace the current wall color with the provided colors.

ABSTRACT

This work focuses on Magic-wall, an automatic system for visualizing the effect of room decoration. Given an image of the indoor scene and a preferred color, the Magic-wall can automatically locate the wall regions in the image and smoothly replace the existing color with the required one. The key idea of the proposed Magic-wall is to leverage visual semantics to guide the entire process of color substitution including wall segmentation and color replacement. We propose an edge-aware fully convolutional neural network (FCN) for indoor semantic scene parsing, in which a novel edge-prior branch is introduced to better identify the boundary of different semantic regions. To accurately localize the wall regions, we adapt a semantic-dependent optimized strategy, which pays more attention to those pixels belonging to the wall by adapting larger optimization weights compared with those from other semantic regions. Finally, to naturally replace the color of original walls, a simple yet effective color space conversion method is proposed for replacement with brightness reservation. We build a new indoor scene dataset upon ADE2 0K [41] for training and testing, which includes 6 semantic labels. Extensive experimental evaluations and visualizations well demonstrate that the proposed Magic-wall is effective and can automatically generate a set of visually pleasing results.

KEYWORDS

Deep Learning; Scene Parsing; Edge Detection

ACM Reference format:
Ting Liu, Yunchao Wei, Yao Zhao, Si Liu, and Shikui Wei. 2017. Magic-wall: Visualizing Room Decoration. In *Proceedings of MM'17, October 23–27, 2017, Mountain View, CA, USA.*, , 9 pages.
DOI: https://doi.org/10.1145/3123266.3123398

*Yunchao Wei is the mentor of this work.
†Shikui Wei is the corresponding author.

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10...$15.00
DOI: https://doi.org/10.1145/3123266.3123398

1 INTRODUCTION

—*Interior design is the art and science of enhancing the interiors of a space or building, to achieve a healthier and more aesthetically pleasing environment for the end user.*[1]

With the development of the society, people gradually pay much more attention to living and working environment. Particularly, the walls' color of living rooms or offices indeed have impact on moods and thoughts of peoples there. For instance, the warm colors (*e.g. red*, *yellow* and *orange*) can spark a variety of emotions ranging from comfort and warmth to hostility and anger, while the cool colors (*e.g. green*, *blue* and *purple*) often spark feelings of calmness as well as sadness. In addition, colors have influences on people in many other ways, depending on the age, gender, ethnic background and occupation. Therefore, it's important to choose wall colors wisely when it comes to decorating. Nowadays, there is a large variety of colors for wall painting. In general, we may easily choose several candidate colors for the target room, according to personal desires or the function of the room. However, it is difficult to determine which color fits best. Thus, our goal is to develop a system to perform automatically wall painting for indoor scene images, so that people can have a look at the room with preferred colored-walls before making the last decision for painting.

Although it seems very desirable, performing automatically indoor wall painting is a challenging task for the following reasons. Firstly, indoor scenes primarily contain a lot of furniture (*e.g. sofa* and *bed*). There exist strong occlusions between furniture and walls. In addition, *windows*, *doors* and items (*e.g. clock* and *photo frame*) hanging on the wall also enhance the difficulty of segmenting the wall regions. Secondly, the edges of the wall are usually hard to be identified, owing to the similarity with other semantic parts of indoor scenes, *e.g. ceiling*. Thirdly, since the distribution of light on the wall is not uniform, it is still difficult to replace the original color with the target color even with satisfactory wall segments.

To address the raised issues, we propose a semantic-aware approach called Magic-wall for wall color editing in this work (see Figure 1). In particular, the Magic-wall is able to automatically locate the wall regions and naturally substitute the current color of the walls with the desirable colors. Magic-wall leverages visual semantics to guide the entire process of wall color editing, including wall segmentation and color replacement. Basically, to effectively generate a dense pixel-wise prediction of semantic labels, we adopt the state-of-the-art semantic segmentation framework, *i.e.* deep Fully Convolutional Neural Network (FCN) [23], as the backbone of the proposed Magic-wall for parsing an input indoor image. In particular, we propose an edge-aware-FCN to better identify the edges of the wall regions, in which a novel edge-prior branch has been introduced to edge prediction. Those features rich in edge information are then utilized for predicting the pixel-level semantics of indoor scene images. In addition, to replace the color of the original wall smoothly and naturally, we present a simple yet effective color space conversion approach for color replacement with brightness reservation. We build a new indoor scene dataset upon the well-annotated large-scale scene parsing dataset ADE20K [41]. Since our target is to segment walls, we only consider the semantics associated with *wall* and re-organize the pixel-level annotations for learning to segment.

[1] https://en.wikipedia.org/wiki/Interior_design

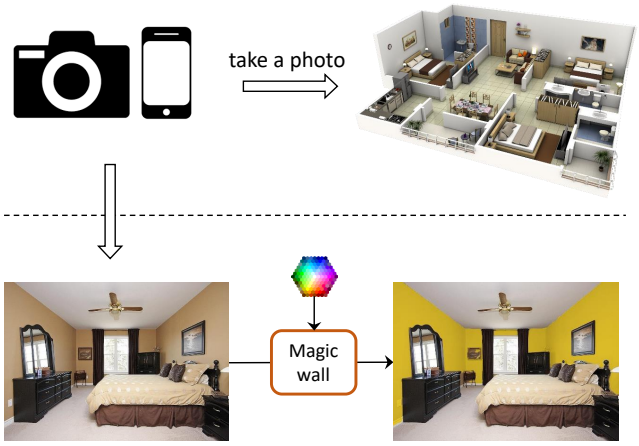

Figure 2: The overall illustration of the proposed Magic-wall system. The user first takes a photo of the target room. The photo accompanying with a chosen color is fed into the Magic-wall system to generate the result.

In summary, the contributions of this work for automatic wall color replacement are as follows:

- We develop an automatic system called Magic-wall for visualizing the effect of room decoration. Extensive visualizations have well demonstrated that the proposed Magic-wall is fully capable to generate a set of visually pleasing results.
- We propose an edge-ware-FCN for effectively learning to semantic segmentation by leveraging the predictive edge information from a novel edge-prior branch. Experimental comparisons well demonstrate the effectiveness of the proposed edge-aware-FCN.
- We propose to employ a simple yet effective color space conversion approach for color replacement with brightness reservation, so that the proposed Magic-wall can produce realism results.

The organization of the rest of this paper is as follows. In Section 2, we provide a review of the related work. Section 3 makes an overview of our proposed Magic-wall. Next, in Section 4, more detailed introductions of the Magic-wall, including edge-aware-FCN and color replacement, are presented. The experimental results are shown later in Section 5. Finally, we conclude this work in Section 6.

2 RELATED WORK

2.1 Indoor scene parsing

There is a rich history of exploration in the field of scene parsing. Liu *et al.* [21] proposed to use label transfer for scene

parsing. Since convolutional neural networks tremendously improved the performance of image classification[1, 15, 30, 31, 38], many CNN-based methods are proposed. In particular, since Long *et al.* [23] proposed fully convolutional network (FCN) that replaced the fully connected layer with convolution layer and achieved remarkable advances in semantic segmentation [8, 35–37], after which many FCN-based approaches have been presented for addressing scene parsing. In later works, Noh *et al.* [25] learned a multi-layer deconvolution network to obtain the coarse-to-fine segmentation. While the segmentation results produced by above-mentioned methods are somewhat coarse, Chen *et al.* [9, 10] proposed the dilated convolution to enlarge receptive field of neural networks that can better localize the object boundaries. In the meantime, a fully-connected CRF was employed as post-processing, which increased the segmentation accuracy near object boundaries significantly. Recently, Zhao *et al.* [40] proposed a pyramid scene parsing network to incorporate global features with local, which achieved excellent performance on scene parsing. In addition, many approaches of indoor scene parsing have been proposed based on RGBD data [11, 26, 28, 29, 33]. Silberman and Fergus [28, 29] addressed this problem by exploring depth information to assist with indoor scene segmentation. Taylor and Cowley [33] also proposed to parse the structure of indoor scene from a RGBD image.

2.2 Edge detection

As a fundamental and important task in computer vision, edge detection has a very long history. Recently, several works has explored convolutional neural networks to detect edge and achieved excellent performance, such as Deep-Contour [27], DeepEdge [4], CSCNN [17], and HED [39]. For instance, Xie and Tu [39] leveraged fully convolutional neural networks and deeply-supervised nets for edge detection. Bertasius *et al.* [5] predicted boundaries by utilizing object-level features, and the boundaries were then applied to facilitate semantic segmentation. Liang *et al.* [20] also proposed to incorporate semantic edge context for human parsing. With the same purpose, Chen *et al.* [8] learned object contours to optimize the semantic segmentation task with a domain transform edge-preserving filter. The main difference between our approach with this work is that they used the edge map to guide the semantic segmentation optimization, while we learn to segment by taking advantages of the intermediate edge-aware convolutional features.

2.3 Appearance assignment and composing

Several studies were made on assigning appearance to a 3D model that can be used to assist users, such as [3, 6, 7, 18, 24]. Nguyen *et al.* [24] proposed a technique to transfer the material style from source images into a target 3D scene. Chen *et al.* [7] presented a system that automatically assign material properties to all objects parts in the 3D scene. The previous works described above have a little parallelism with ours. Nevertheless, their algorithms are developed primarily to 3D scene material suggestions. Besides, the core problem of their works is to define the material and aesthetic rules that can be solved by combinatorial optimization.

The main purpose of appearance composing is to seamlessly combine the given images. Tao *et al.* [32] presented an algorithm for minimizing artifacts in gradient-domain image compositing. Tsai *et al.* [34] proposed an automatic background replacement algorithm that generated images with diverse stylized skies. Liu *et al.* [22] proposed to automatically synthesis the makeup for a female's face by a novel Deep Localized Makeup Transfer Network.

However, to generate more realistic results by Magic-wall, it is necessary to meet with the high requirements for well matching between images under-merged. Thus, as the approaches mentioned above, one of the important steps is searching a set of images similar to input image from the dataset. Instead, users can choose any color they like for replacement in our system.

3 OVERVIEW

Given an input indoor scene image, we aim to automatically generate a set of results, in which the color of walls is naturally replaced. To achieve this, the key problem is how to successfully separate the wall regions with others. In this paper, we propose an edge-aware-FCN for indoor scene parsing. Figure 3 shows the overview of the framework, which is the critical component of the Magic-wall system. The network is built upon the Deeplab-LargeFOV [10], whose parameters are initialized by VGG16 pre-trained on ImageNet [13]. In general, Convolutional Neural Networks (CNNs) usually use several hidden layers to hierarchically learn high-level representation of images. In this case, the layers at the front (end) of networks usually perceive the low-level features (high-level semantics) of the input image. As shown in Figure 3, we predict edges by leveraging features from the front several layers, *i.e. conv4* and *conv5*, and produce the dense pixel-wise prediction of semantic labels using the last convolutional layer, *i.e. conv7*. In particular,

- **Edge prediction** We conduct convolutional operations on the last feature maps of each convolutional group (from *conv4* and *conv5*), as shown in Figure 3. The produced convolutional feature maps are reduced to 2-channel confidence maps, which are further concatenated for edge prediction.
- **Pixel-wise semantic prediction** To enhance the capability of semantic representation for the last convolutional layer, we adopt the convolutional feature map from *conv7* to predict dense pixel-level semantic labels. Meanwhile, we also exploit edge-aware feature maps for better semantic segmentation. In particular, the features produced by the intermediate layers from edge-ware branch, as shown in Figure 3, are concatenated with that from *conv7* for making the dense pixel-wise prediction of semantic labels.

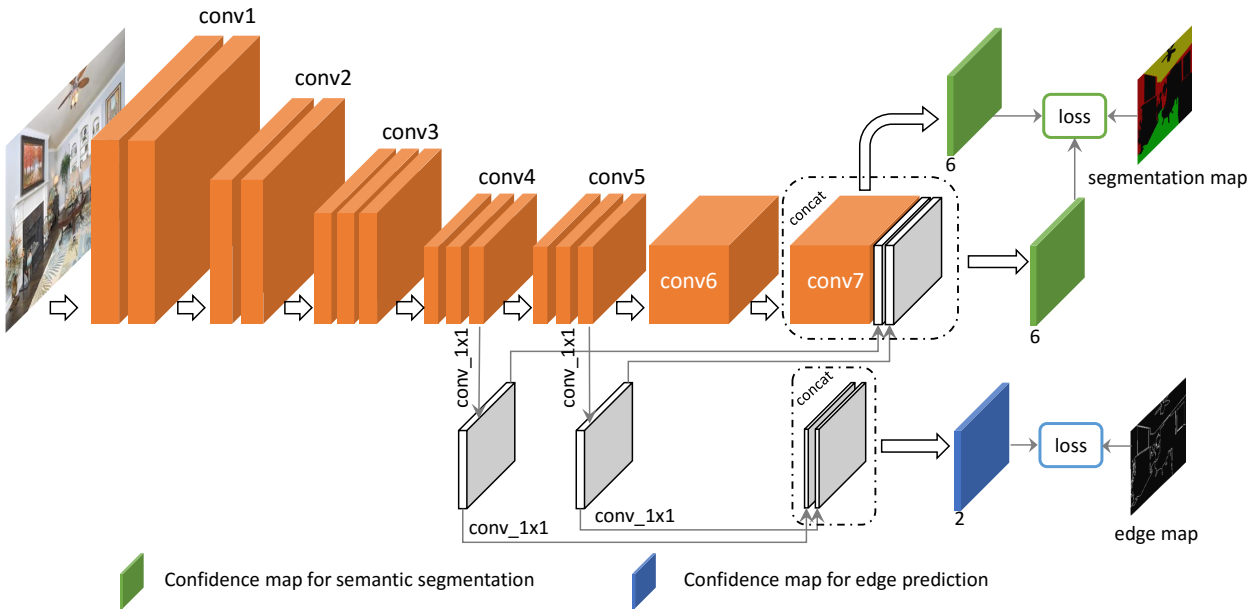

Figure 3: Overview of the proposed edge-aware-FCN for semantic segmentation.

We resize the feature maps to gain the same spatial resolution by bilinear interpolation before concatenation and conduct the convolutional operation with 1×1 kernel size to produce semantic labels and edge predictions. Based on the obtained wall regions, we conduct the replacement operation in the HSV color space for keeping brightness information, which is able to strengthen the reality of the produced indoor scene image.

4 MAGIC-WALL MODEL

In this section, we first explain the proposed edge-aware-FCN more formally. Then, the details of color replacement with brightness reservation are introduced in the following.

4.1 Edge-aware-FCN

Suppose the training set $\mathcal{I}$ includes N images. We denote I as any image from $\mathcal{I}$ and S is the corresponding pixel-wise segmentation mask. Based on S, we further produce the edge mask E for I, in which the pixels between two semantic regions are set as 1 and others are set as 0. Denote the sets of semantic labels and edge labels for segmentation and edge prediction as $\mathcal{C}_{seg}$ and $\mathcal{C}_{edge}$, respectively. We assume that all the semantic labels are included in each training image. Our target is to train a segmentation network $f(I;\theta)$ parameterized by θ, which predicts the pixel-wise probability of each label $c \in \mathcal{C}_{seg}$ or $c \in \mathcal{C}_{edge}$ at each location u of the image plane $f_{u,c}(I;\theta)$. Totally, the cross-entropy loss used for optimizing the edge-aware-FCN is formulated as

$$\min_{\theta} \sum_{I \in \mathcal{I}} L_{edge}(f(I;\theta)) + L_{seg}(f(I;\theta)), \tag{1}$$

where $L_{edge}(f(I;\theta))$ and $L_{seg}(f(I;\theta))$ are the loss functions for edge prediction and pixel-wise semantic segmentation, respectively.

Edge Prediction We consider that the edge-aware feature map generates from the $i^{th}(i = 4, 5)$ convolutional feature map with 1×1 convolutional filters as F_{e_i}, which is further employed to produce confidence map for edge prediction. To increase the edge prediction accuracy, we first concatenate the confidence maps from multi-level prediction streams as show in Figure 3, and then a convolutional operation with 1×1 kernel size is conducted to fuse the individual confidence maps into the final one, *i.e.* P_e.

Based on the generated P_e, the loss function for predicting edge map of I can be formulated as

$$L_{edge} = -\frac{1}{\sum_{c \in \mathcal{C}_{edge}} |E_c|} \sum_{c \in \mathcal{C}_{edge}} \sum_{u \in E_c} \log f_{u,c}(P_e;\theta). \tag{2}$$

L_{edge} is computed over all pixels in I; however, over 90% of the pixels do not belong to edges. Following the class-balancing cross-entropy loss advised by [39], L_{edge} can be computed by weighting non-edge and edge pixels with different ratios. We denote λ_0 and λ_1 as the weights of no-edge (indicated by 0) and edge (indicated by 1) pixels. Eqn (2) is then formulated as

$$\begin{aligned} L_{edge} = &-\frac{1}{|E_0|} \sum_{u \in E_0} \lambda_0 \log f_{u,c}(P_e;\theta) \\ &-\frac{1}{|E_1|} \sum_{u \in E_1} \lambda_1 \log f_{u,c}(P_e;\theta). \end{aligned} \tag{3}$$

Semantic Segmentation As shown in Figure 3, we employ two optimization terms for making the dense pixel-wise prediction, including one non-edge-aware term (L^{ne}_{seg}) and one

edge-aware term (L^e_{seg}). In general, we formulate the loss function of the pixel-level semantic segmentation as

$$L_{seg} = L^{ne}_{seg} + L^{e}_{seg}. \tag{4}$$

We employ the L^{ne}_{seg} to encourage the feature maps from *conv7* (denoted as F_{c_7}) to perceive high-level semantics of indoor scene. The corresponding confidence map P^{ne}_{seg} is then produced by conducting convolutional operation with kernel of 1×1 size on F_{c_7}. Meanwhile, to leverage edge-aware feature maps for improving the quality of predicted segmentation confidence map, we integrate F_{c_7} with F_{e_i} from the edge prediction branch with a concatenated operation to construct the edge-aware feature representation (denoted as F_{c_7+e}). With the confidence map P^{e}_{seg} calculated from F_{c_7+e}, the loss function for predicting segmentation mask of I can be formulated as

$$\begin{aligned} L_{seg} = &- \frac{1}{\sum_{c\in\mathcal{C}_{seg}} |S_c|} \sum_{c\in\mathcal{C}_{seg}} \sum_{u\in S_c} \log f_{u,c}(P^{ne}_{seg};\theta) \\ &- \frac{1}{\sum_{c\in\mathcal{C}_{seg}} |S_c|} \sum_{c\in\mathcal{C}_{seg}} \sum_{u\in S_c} \log f_{u,c}(P^{e}_{seg};\theta), \\ \Longrightarrow& \\ L_{seg} = &- \frac{1}{\sum_{c\in\mathcal{C}_{seg}} |S_c|} \sum_{c\in\mathcal{C}_{seg}} \sum_{u\in S_c} \log f_{u,c}(P^{ne}_{seg}, P^{e}_{seg};\theta). \end{aligned} \tag{5}$$

Since our target is to accurately locate pixels belonging to the wall for color replacement, we prefer the edge-aware-FCN to produce the segmentation results more accurate on the *wall* semantics. To achieve this goal, we adopt a larger optimized weight for the pixels belonging to *wall* compared with those of other semantics. Formally, suppose the weights for wall-pixel and non-wall-pixel as η_w and η_{nw}, Eqn (5) can be re-formulated as

$$\begin{aligned} L_{seg} = &- \frac{1}{|S_{wall}|} \sum_{u\in S_{wall}} \eta_w \log f_{u,c}(P^{ne}_{seg}, P^{e}_{seg};\theta) \\ &- \frac{1}{\sum_{c\neq wall} |S_c|} \sum_{c\neq wall} \sum_{u\in S_c} \eta_{nw} \log f_{u,c}(P^{ne}_{seg}, P^{e}_{seg};\theta). \end{aligned} \tag{6}$$

All the parameter settings are detailed in Section 5. With the learned edge-aware-FCN, the semantic segmentation mask of the testing image can be obtained. We then extract the regions (denoted as W) belonging to *wall* for the following color replacement.

4.2 Color Replacement

Information of brightness and shadow is highly demanded by natural replaced results, so we need a color space that brightness channel is independent from other channels. As we know, HSV, also called HSB (in which B for brightness), stands for hue, saturation and value, which just meets our requirements. Therefore, the entire replacement process is implemented in HSV color space. However, results of the segmentation around the boundaries, which is directly obtained from the network, are coarse. Thus, a global sampling method [14] for alpha matting is employed to refine the segmentation results.

Given an input image I, we first extract the wall mask W and convert the image from RGB to HSV color space, where W is a binary mask. If pixel in position x belongs to the wall, $W^x = 1$; otherwise, $W^x = 0$. To ensure the edge smooth, we generate the trimap according the binary mask W. Specifically, the pixels around the edge of the wall are set as unknown pixels, while the others are set as known foreground/background. With the global sampling method, alpha matte α can be computed, where α^x is the opacity of the wall in pixel x. Finally, the following formulation is used to guide the entire replacement procedure:

$$\begin{aligned} h^x_O &= \alpha^x h_R + (1-\alpha^x)\, h^x_I \\ s^x_O &= \alpha^x s_R + (1-\alpha^x)\, s^x_I \\ v^x_O &= (\alpha^x - \beta\alpha^x)\, v_R + (1-\alpha^x+\beta\alpha^x)\, v^x_I \quad \beta\in[0,1], \end{aligned} \tag{7}$$

where h_O, h_I and h_R represent hue of output image, input image and reference color respectively. Analogically, s and v denote saturation and value, and β decides the extent of value(V) that the input image reserved. If β is too small, the output image will loss brightness and shadow information. Conversely, it will tarnish the reference color.

5 EXPERIMENTAL RESULTS

5.1 Dataset and Settings

Dataset We evaluate our proposed approach on a dataset built upon ADE20K [41]. ADE20K is a dataset of indoor scenes used in MIT scene parsing challenge 2016, which has a total of 25K images of 150 classes with a variety of semantics about indoor scenes. For our specific task, only the images with the wall appeared are required. Therefore, the images without the wall are removed. And then, to train the FCN network, four semantic labels frequently accompanied with the wall are selected, including floor, ceiling, window, and table. The remaining labels are set as background. Finally, 3000 images are selected, in which 2500 and 500 images are used for training and testing, respectively.

Training/Testing Settings We utilize DeepLab code [10] in our experiments, which is implemented based on the publicly available Caffe framework [19]. Standard stochastic gradient descent is employed for optimization, where initial learning rate, momentum and weight decay are set to 0.001, 0.9 and 0.0005, respectively. Following [10], we adapt the "poly" learning rate policy where the learning rate is changed every iteration by multiplying a factor of $(1-\frac{iter}{maxiter})^{power}$. All the newly added layers are randomly initialized with zero-mean Gaussian distributions with standard deviations of 0.01. Each intermediate convolutional layers of edge branches has 128 kernels of size 1×1. The iteration number is set to 10K with batch size of 4. Due to large cropsize can get good performance, the training images is cropped to 473×473. The whole networks are initialized with the pre-trained VGG-16 model provided by [10]. For comparison, we adopt Deeplab-LargeFOV as the baseline model.

Table 1: The comparison of mIoU(%) for baseline and our proposed edge-aware-FCN. NE denotes non-edge-aware loss.

Method	Background	Wall	Floor	Ceiling	Window	Table	mIoU
Deeplab-LargeFOV	72.125	72.464	68.794	79.031	54.593	36.794	63.967
Edge-aware-FCN	73.629	74.501	**69.929**	81.079	56.342	37.468	65.492
Edge-aware-FCN+NE	**73.908**	**75.196**	69.518	**81.961**	**56.391**	**37.805**	**65.796**

Evaluation Metrics As region intersection over union (IoU) takes into account both the false and the missed values for each class, it has been a standard semantic segmentation evaluation. Thus, we employ the IoU for semantic segmentation evaluation in our experiments. Besides, to evaluate the replacement quality, we ask 21 participants to conduct the user study according to the verisimilitude of the composed images. The participants rate the results into five degrees: "excellent", "good", "average", "fair", and "poor". At last, we count the percentage of every degree to report the performance of our proposed approach.

5.2 Ablation Analysis

5.2.1 Edge-aware-FCN. We adopt Deeplab-LargeFOV as the baseline model, and we report the performance of our Edge-aware-FCN comparing with baseline method in Table 1. All the experiments are conducted with same experiment settings. Since we find that the IoU on the wall drops near 1% without non-edge-aware loss, non-edge-aware loss is added in the following experiments. As shown in Table 1, it is obvious that our proposed method outperforms the baseline model significantly. We observe that the mIoU score is improved from 63.967% to 65.796%. Besides, the mIoU on the wall is increased from 72.464% to 75.196%, with 2.73% improvement. It demonstrates that introducing intermediate edge features indeed promote the performance of semantic segmentation. Several visualized results are shown in Figure 4 for demonstrating clearly. From the region in the white box, we can see that our method performs better in capturing the object counters than baseline, which indicating that our model really captured the useful edge features and improved the semantic segmentation. In addition, we also display the semantic edge predicted by edge branches. Note that we aim at predicting semantic edge to promote the semantic segmentation, while the evaluation of the precision for edge is not involved.

We now proceed to evaluate the edge-aware branch and investigate how it benefits the semantic segmentation. The parameters λ_0 and λ_1 in Eqn (3) can be computed by the ratio of non-edge and edge pixels in the image. Based on the baseline model, we add two edge prediction branches to the front convolutional layers (*conv4* and *conv5*), and the middle intermediate convolutional layers are concatenated with *conv7* together to facilitate the semantic segmentation. To find out how to structure the edge-aware branch, we investigate the effect by varying convolutional layers combination for edge detection. Note that our network aims at predict semantic edge, the boundaries between different semantic regions in the mask, which is not correspond to traditional edge detection. Moreover, the layers at the front of networks always capture more low-level features, the later capture more high-level semantic features. Therefore, the edge detection structure begins with *conv5* only and is varied by adding previous layers gradually. As shown in Table 2, all the performances are improved compared with that of Deeplab-LargeFOV, which indicates the effectiveness of edge branches. We observe that the mIoU drops by near 1% when only uses *conv5* layer or all convolutional layers compared with other combinations. The reason is that the front layers capture more low-level features that do not corresponding to semantic edge, while more detailed information is lost with only *conv5* layer. Thus, according to the Table 2, we perform the edge prediction from *conv4* and *conv5* for better segmentation performance.

Table 2: The comparison of different combinations for edge detection.

Method	wall(%)	mIoU(%)
Deeplab-LargeFOV	72.464	63.967
conv5	74.476	64.913
conv4+conv5	**75.196**	**65.796**
conv3+conv4+conv5	74.967	65.443
conv2+conv3+conv4+conv5	74.543	65.197
conv1+conv2+conv3+conv4+conv5	74.188	64.701

To further enhance the segmentation performance on the wall, we want the network to pay more attention to the *wall* semantic. Therefore, weighted cross-entropy loss is introduced in the final network, which results in 0.5 improvement on the wall. The weight on the wall is set as 1.5 empirically by the validation, and all others are set as 1. After employing CRF for post-processing, the IoU on the wall finally achieves 76.534%. The gradually improved performance validates the effectiveness of the proposed edge-aware-FCN.

5.2.2 Color Space Conversion. Figure 5 evaluates the effects of segmentation refinement and brightness reservation. We first compare the results with and without segmentation refinement. As shown in Figure 5(b), without segmentation refinement, the edges of the wall are not smooth, and the table lamp in the red box is segmented imprecisely. Observing Figure 5(c), the edge between the wall and the ceiling is smooth and natural with segmentation refinement. In addition, the contour of the table lamp is more accurate than

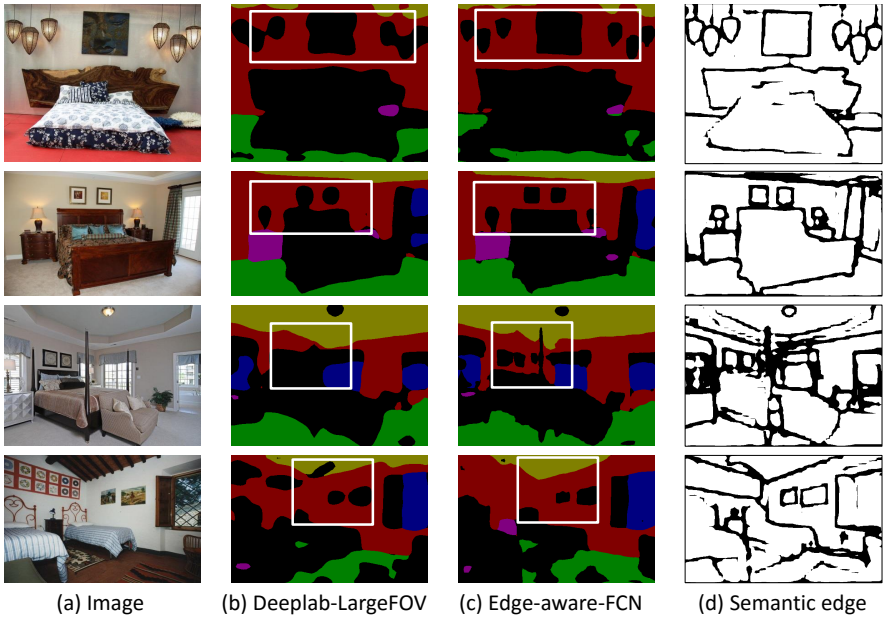

Figure 4: Visualized semantic segmentation results. (a) The input image. (b) Scene parsing results of baseline network. (c) Scene parsing results of our Edge-aware-FCN. (d) Edge detection results of our Edge-aware-FCN.

Figure 5: Examples of replacement result with and without refinement and brightness reservation.

Figure 5(b). Obviously, Figure 5(b) looks better than Figure 5(c). That's because the non-edge pixels, which locate around the boundaries of the wall, are assigned low alpha value after processed by alpha matting. Thus, with the replacement under the guide of the alpha matte, the edges around the wall look more smooth.

We now discuss the importance of color space conversion. To balance the illumination between input image and reference color, the weight β in Eqn (7) is set as 0.5. As shown in Figure 5(c), without brightness reservation, Comparing Figure 5(d) with Figure 5(c), we can see that the glow of the table lamp in Figure 5(d) is reserved, which makes the replaced regions look more harmonious and verisimilar. The reason is that the Figure 5(d) keeps the same illumination and shadow information with input image by reserving the original brightness, which can generate a more realistic result.

In order to measure the quality of the proposed replacement algorithm, we conduct a user study to evaluate the verisimilitude of generated images. The baseline is directly replacing the wall in RGB color space with the segmentation mask obtained from networks. Now we mainly compare the baseline with the proposed segmentation refinement and brightness reservation. We randomly select 50 images from dataset. Each time, a 3-tuple generated by the three methods is sent to 21 participants to mark. Note that the three images in every tuple are shown randomly. We report the percentages of every degree in Table 3. It can be found that the scores of baseline are mainly distributed in "fair". Instead, segmentation refinement(SR) is concentrated on "average". It indicates that the replacement with segmentation refinement can generate more harmonious results. With the brightness reserving(BR), the scores are mainly distributed in "good". Furthermore, 65.9% of the results are marked as "good" even "excellent", which is much higher than 11.43% of the baseline. Therefore, our algorithm can produce more realistic results, which demonstrates the significant advantages of the proposed approach.

To further validate the effectiveness of Magic-wall, we also conduct additional experiments of texture replacement.

Figure 6: Examples of indoor images with the color and texture of the wall replaced. The first column shows the input images. In each group, the first line shows the results with four colors, and the second line shows results with four textures.

Table 3: The comparison of quality for different methods. SR denotes segmentation refinement, and BR denotes brightness reserving.

Method	excellent	good	average	fair	poor
Baseline	0.86%	10.57%	25.90%	41.05%	21.62%
SR	3.14%	22.29%	40.19%	24.67%	9.71%
SR+BR	18.76%	47.14%	25.14%	7.90%	1.05%

Some successful examples are shown together with color replacement in Figure 6 (some input images are taken from Internet). In spite of the excellent generated results, the texture replacement still has some problems. In fact, there should have a warp when an image is taken from a 3D space. As shown in Figure 6, we can see that the image lost the sense of space in the corner of the wall, which makes the results a little unrealistic. In the future, we can leverage layout estimation [2, 12, 16] and homography matrix to warp the texture image for more naturalistic effect.

6 CONCLUSION

In this paper, we propose an edge-aware fully convolutional neural network and a color space conversion method to automatically replace the current color of the wall with the provided color. The proposed edge-aware-FCN further improves the segmentation comparing with deeplab. To smooth the edge of the wall and generate more realistic results, we utilize alpha matting to refine the segmentation, and the color space conversion is employed to perform the replacement procedure. Extensive experiments and examples show that our system can compose more realistic images. In the future work, we intend to combine semantic segmentation with layout estimation for more authentic results.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (No.61532005, No.61572065, No.61572493, No.U1536203), National Key Research and Development of China (No.2016YFB08004 04), Joint Fund of Ministry of Education of China and China Mobile (No.MCM20160102),and the Fundamental Research Funds for the Central Universities (No.2017YJS055).

REFERENCES

[1] K. Alex, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.

[2] M. Arun and L. Svetlana. 2015. Learning informative edge maps for indoor scene layout prediction. In *ICCV*. 936–944.

[3] S. Bell, P. Upchurch, N. Snavely, and K. Bala. 2013. OpenSurfaces: A Richly Annotated Catalog of Surface Appearance. *ACM Trans. Graph.* (2013), 111:1–111:17.

[4] G. Bertasius, J. Shi, and L. Torresani. 2015. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *CVPR*. 4380–4389.

[5] G. Bertasius, J. Shi, and L. Torresani. 2015. High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *ICCV*. 504–512.

[6] M. G. Chajdas, S. Lefebvre, and M. Stamminger. 2010. Assisted Texture Assignment. In *i3D*.

[7] K. Chen, K. Xu, Y. Yu, T. Wang, and S. Hu. 2015. Magic decorator: automatic material suggestion for indoor digital scenes. *TOG* 34, 6 (2015), 232.

[8] L. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. 2016. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*. 4545–4554.

[9] L. Chen, P. George, K. Iasonas, M. Kevin, and A. L. Yuille. 2015. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*.

[10] L. Chen, P. George, K. Iasonas, M. Kevin, and A. L. Yuille. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv:1606.00915* (2016).

[11] C. Couprie, C. Farabet, L. Najman, and Y. Lecun. 2013. Indoor semantic segmentation using depth information. In *ICLR*.

[12] S. Dasgupta, K. Fang, K. Chen, and S. Savarese. 2016. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *CVPR*. 616–624.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

[14] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun. 2011. A global sampling method for alpha matting. In *CVPR*. 2049–2056.

[15] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[16] V. Hedau, D. Hoiem, and D. Forsyth. 2009. Recovering the Spatial Layout of Cluttered Rooms. In *ICCV*.

[17] J. Hwang and T. Liu. 2015. Pixel-wise deep learning for contour detection. *arXiv preprint arXiv:1504.01989* (2015).

[18] A. Jain, T. Thormählen, T. Ritschel, and H. P. Seidel. 2012. Material Memex: Automatic Material Suggestions for 3D Objects. *TOG* 31, 5 (2012).

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[20] X Liang, C Xu, X Shen, J Yang, J Tang, L Lin, and S Yan. 2016. Human Parsing with Contextualized Convolutional Neural Network. *PAMI* (2016).

[21] C. Liu, Y. Jenny, and A. Torralba. 2011. Nonparametric scene parsing via label transfer. *PAMI* 33, 12 (2011), 2368–2382.

[22] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao. 2016. Makeup like a superstar: Deep Localized Makeup Transfer Network. *arXiv preprint arXiv:1604.07102* (2016).

[23] J. Long, E. Shelhamer, and T. Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. 3431–3440.

[24] C. H. Nguyen, T. Ritschel, K. Myszkowski, E. Eisemann, and H. Seidel. 2012. 3D material style transfer. In *CGF*, Vol. 31. 431–438.

[25] H. Noh, S. Hong, and B. Han. 2015. Learning deconvolution network for semantic segmentation. In *ICCV*. 1520–1528.

[26] X. Ren, L. Bo, and D. Fox. 2012. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*. 2759–2766.

[27] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. 2015. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *CVPR*. 3982–3991.

[28] N. Silberman and R. Fergus. 2011. Indoor scene segmentation using a structured light sensor. In *ICCV*. 601–608.

[29] N. Silberman, D. Hoiem, P. Kohlit, and R. Fergus. 2012. Indoor segmentation and support inference from rgbd images. *ECCV* (2012), 746–760.

[30] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9.

[32] M. W. Tao, M. K. Johnson, and S. Paris. 2010. Error-tolerant Image Compositing. In *ECCV*.

[33] C. J. Taylor and A. Cowley. 2013. Parsing indoor scenes using rgb-d imagery. In *Robotics: Science and Systems*, Vol. 8. 401–408.

[34] Y. Tsai, X. Shen, Z. Lin, K. Sunkavalli, and M. Yang. 2016. Sky is not the limit: Semantic-aware sky replacement. *TOG* 35, 4 (2016), 149.

[35] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan. 2017. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *CVPR*.

[36] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, and S. Yan. 2016. Learning to segment with image-level annotations. *Pattern Recognition* 59 (2016), 234–244.

[37] Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, J. Feng, Y. Zhao, and S. Yan. 2016. STC: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI* (2016).

[38] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. 2016. HCP: A flexible CNN framework for multi-label image classification. *TPAMI* 38, 9 (2016), 1901–1907.

[39] S. Xie and Z. Tu. 2015. Holistically-nested edge detection. In *ICCV*. 1395–1403.

[40] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. 2016. Pyramid Scene Parsing Network. *CoRR* abs/1612.01105 (2016).

[41] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. 2016. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442* (2016).