Magic-wall: Visualizing Room Decoration by Enhanced Wall Segmentation

Ting Liu, Yunchao Wei, Yao Zhao, *Senior Member, IEEE*, Si Liu, Shikui Wei

***Abstract*—This work presents an intelligent system named Magic-wall, which enables to visualize the effect of room decoration automatically. Concretely, 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 wall 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 replacement. To strengthen the reality of visualization, we make the following contributions. First, we propose an edge-aware fully convolutional neural network (Edge-aware-FCN) for indoor semantic scene parsing, in which a novel edge-prior branch is introduced to identify the boundary of different semantic regions better. To further polish the details between the wall and other semantic regions, we leverage the output of Edge-aware-FCN as the prior knowledge, concatenating with the image to form a new input for the Enhanced-Net. In such a case, the Enhanced-Net is able to capture more semantic-aware information from input and polish some ambiguous regions. Finally, to naturally replace the color of the original walls, a simple yet effective color space conversion method is proposed for replacement with brightness reserved. We build a new indoor scene dataset upon ADE20K [1] 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.**

***Index Terms*—Deep Learning, Scene Parsing, Edge Detection**

I. Introduction

In the interior decoration, the color of the wall painting is crucial to the final effect. It can accentuate existing architectural details, add interest to a room as well as reflect people's personality directly. A color sets the mood for a room's interior and conveys how people want the space to feel. Usually, light colors like green, sunshine yellow, and tangerine are expansive and airy, making rooms seem larger and brighter. While dark colors like red, blue and brown are sophisticated and warm; they give large rooms a more intimate appearance. 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 wall painting automatically 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.

To achieve this goal, we propose a semantic-aware approach called Magic-wall for wall color editing in this work. As shown in Fig. 1, the user first takes a photo from the target room, and then the indoor scene image with a preferred color are fed into our Magic-wall system to generate the final visual effect. 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 color. 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) [2], [3], as the backbone of the proposed Magic-wall for parsing an input indoor image.

Designing such an automatic wall painting system is non-trivial. Unfortunately, performing automatically indoor wall painting is very challenging as there are three issues need to be addressed. First, the edges of the wall are usually hard to be identified. Due to indoor scenes primarily contain a lot of furniture, strong occlusions usually appear between furniture and walls. In addition, the similarity with other semantic parts of indoor scenes makes it more difficult as well, *e.g. ceiling*. Second, there usually exist blurred regions which are hard to be localized. For instance, the little items (*e.g. clock*, *photo frame* and *switch*) hanging on the walls, especially the leaves of the potted plant around the walls. Third, how to naturally and smoothly replace the color need to be solved. 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. At last, the authentic texture replacement still is a challenge. To generate a natural effect, the texture should be deformed in the corner. However, it's hard to perform the deformation precisely.

To address the raised issues, we make the following contributions to gradually improve the quality of details as well as the accuracy of the predicted segmentation masks. Firstly, we propose an Edge-aware-FCN to identify better the edges of the wall regions, in which a novel edge-prior branch has

This work was supported in part by National Key Research and Development of China (No.2016YFB0800404), National Natural Science Foundation of China (No.61532005, No.61572065), Program of China Scholarships Council (No.201807095006), Fundamental Research Funds for the Central Universities (No. 2018JBZ001). (*Corresponding author: Yunchao Wei*)

T. Liu, Y. Zhao, and S. Wei are with Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: 16112055@bjtu.edu.cn; yzhao@bjtu.edu.cn; shkwei@bjtu.edu.cn;).

Y. Wei is with Beckman Institute, University of Illinois at Urbana-Champaign (e-mail: wychao1987@gmail.com).

S. Liu is with Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: liusi@buaa.edu.cn).

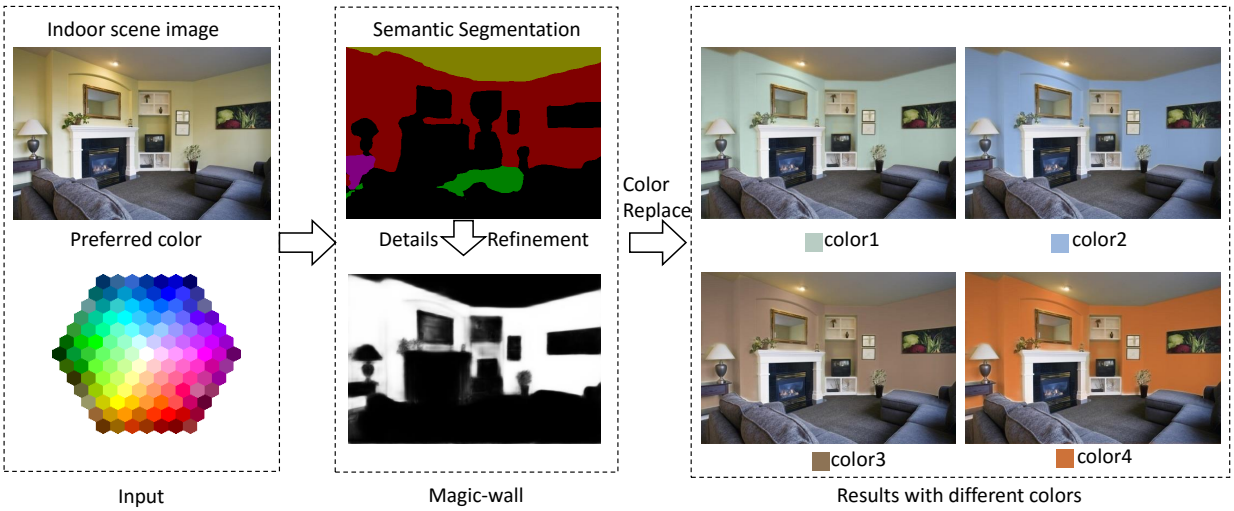


Fig. 1. The overall illustration of the proposed Magic-wall system. The user first takes a photo of the target room. The photo accompanying with a chosen color is fed into the Magic-wall system to generate the result. The demo is available at http://mic.bjtu.edu.cn/project/wall_demo.

been introduced to edge prediction. Those features rich in edge information are then utilized for predicting the pixel-level semantics of indoor scene images. Secondly, inspired by [4], [5], we introduce an Enhanced-Net to further enhance the segmentation quality of wall regions. Nevertheless, the image firstly goes through the Edge-aware-FCN. The produced semantic confidence map is then concatenated with the RGB image to form into a new input for the Enhanced-Net. In such a case, the prior semantic-aware knowledge mined by the Edge-aware-FCN can be leveraged by Enhanced-Net and the Enhanced-Net is encouraged to focus on polishing more challenging and ambiguous details around wall regions. We assemble Edge-aware-FCN and Enhanced-Net into a unified framework, which can be learned in an end to end manner. Finally, 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 reserved. For texture replacement, the transformation can be solved by utilizing the indoor scene layout estimation. However, the existing indoor scene layout estimation algorithms still can't meet our requirement of highly precise estimation. Thus, we consider it as the future work. We build a new indoor scene dataset upon the well-annotated large-scale scene parsing dataset ADE20K [1]. 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 segmentation.

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 of generating a set of visually pleasing results.
- We propose an Edge-aware-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 an Enhanced-Net to further polish the details between the wall and other semantic regions by leveraging the output of Edge-aware-FCN as the prior knowledge and concatenating with the RGB image to form a new input for the Enhanced-Net. Experimental comparisons demonstrate that the Enhanced-Net is able to polish more challenging and ambiguous details around wall regions.
- We propose to employ a simple yet effective color space conversion approach for color replacement with brightness reserved, so that the proposed Magic-wall can produce realism results.

The Magic-wall system we present in this paper has several improvements compared to that reported in our conference version (Magic-wall-V1) [6]: (1) A new post-processing approach called Enhanced-Net is proposed in the updated version for better polishing the details around the wall instead of the time-consuming CRF [7] and Global Matting [8] used in [6]. (2) More extensive experiments have been conducted based on ResNet-101 [9] to further demonstrate the effectiveness of the Edge-aware-FCN and Enhanced-Net. Besides, several training tricks are applied to further improve the performance. (3) The visual effect of the updated version is more authentic as we apply the probability map for wall predicted by Enhanced-Net, in which the details of the wall are polished, to guide the entire substitution procedure. To validate the superiority of the updated version, we also conduct user study according to the verisimilitude of the composed results.

The organization of the rest of this paper is as follows. In Section II, we provide a review of the related work. Section III makes an overview of our proposed Magic-wall. Next, in Section IV, more detailed introductions of the Magic-wall, including Edge-aware-FCN, Enhanced-Net and color replacement, are presented. The experimental results are shown later in Section V. Finally, we conclude this work in Section VI.

II. Related Work

A. Semantic Segmentation

In recent years, semantic segmentation has riveted more and more attention in the field of computer vision. Since deep neural network [10], [11], [12], [9], [13], [14], [15], [16], [17], [18], [19] made a tremendous breakthrough in the image classification task, it also brought dramatic changes for other computer vision task, such as object detection [20], [21], [22], [23] and semantic segmentation [2], [3]. Long *et al.* [2] proposed FCN for pixel-wise classification by replacing the fully connected layers with convolutional layers. Although FCN achieved great improvements, the produced segmentation maps were still too coarse because the consecutive pooling operations discarded detailed information. Following FCN, several improved approaches were presented to generate finer results. Generally, there are the following three approaches to recover the detailed information.

The first method is skip-connections, which has been demonstrated effectively in [2]. It's applied in several works such as ParseNet [24] and RefineNet [25]. Specifically, this type of connection skips multiple layers to link the lower layers and the higher layers. In this manner, the high-resolution features with more details from low layer are incorporated together with the high-level semantic information for finer segmentation.

The second method is applying upsampling operations like deconvolution and unpooling, which is adopted by DeconvNet [26], SegNet [27], ENet [28], U-Net [29], FC-DenseNet [30] and so on. Usually, this kind of network has symmetric architecture, named as encoder-decoder. The encoder network aims at learning rich semantic but low-resolution representations. The decoder is usually resembled with encoder and attempts to compensate the lost details by exploiting the deconvolution and unpooling.

The third method is utilizing dilated convolution [3], [31], [32], like DeeplabV3 [33], PSPNet [34]. In this method, the downsampling operations are skipped in the last several layers to generate the high-resolution feature maps. To keep the same receptive field, dilated convolution operations are performed to enlarge the receptive field of neural networks without increasing the number of parameters.

In fact, the three methods are always integrated together to the best advantage. For instance, some works adopted encoder-decoder architecture and skip connections to mitigate the loss of detailed information like ParseNet [24], ENet [28], U-Net [29] RefineNet [25], FC-DenseNet [30]. By extending DeeplabV3, DeeplabV3+ [35] added a skip connection from the front layer to refine the object boundaries, called decoder in [35]. As an effective postprocessing method, CRF [7] was widely employed to refine the boundaries. To our specific application, we focus on the detailed information and the boundaries of the wall. Hence, we propose the Edge-aware-FCN and Enhanced-Net for precisely segmenting the wall region.

B. Edge detection

Edge detection, which is among the fundamental problems in computer vision, aims to extract the edges from images. It has quite a long history. Recently, a series of works explored convolutional neural networks to detect edge and achieved excellent performance, such as N^4-Fields [36], DeepEdge [37], CSCNN [38], and HED [39]. For instance, Xie and Tu [39] leveraged fully convolutional neural networks and deeply-supervised nets for edge detection. Liu *et al.* [40] achieved improvement by utilizing relaxed label to guide the predictions. Recently, Liu *et al.* [41] exploited multiscale and multilevel information of objects to perform the edge prediction by combining all the meaningful convolutional features in a holistic manner.

Since the edge information with fundamental and structured characteristics is greatly beneficial for the high-level task, it is widely utilized to facilitate other tasks like semantic segmentation [42], [43], [44], [45] and road detection [46], [47], [48]. Bertasius *et al.* [42] and Kokkinos [43] both explored the potential of boundary information for semantic segmentation. Although many previous works proposed to exploit the edge information to improve the semantic segmentation, most of them directly took the edge maps as features. For example, Liang *et al.* [45] integrated learned semantic edge maps into the parsing features to promote the human parsing. Chen *et al.* [44] leveraged the learned object contours as the reference to optimize the semantic segmentation task by a domain transform edge-preserving filtering method. However, we utilize the intermediate edge-aware features (not semantic edge maps) as complementary characteristics for better inferring regions from different semantics. Besides, Wang *et al.* [47] designed a weight-shared siamesed network to compute the features of the input contour map for improving the road detection. Comparing with [47], we both use the edge-aware features to improve the segmentation. However, we obtained the edge-aware features by learning to predict the edge map while [47] generates the edge-aware features on given edge map. In addition, the edge map is no more required during the test phrase for our method.

C. Semantic Stylization

Semantic stylization mainly focuses on simulating images by semantic-driven techniques. A vast majority of publications in the field of semantic stylization have been proposed. For example, Tsai *et al.* [49] developed a system for generating images with diverse stylized skies by an automatic background replacement algorithm. Shen *et al.* [50] transferred the style for facial images with a fully automatic portrait segmentation technique. Liu *et al.* [51] automatically synthesized the makeup for a female's face by a novel Deep Localized Makeup Transfer Network. To enhance the visual effect of image compositing, Tao *et al.* [52] attempted to improve the gradient-domain compositing method. Generally, to produce a realistic effect, the transformed style was decided by the reference images, which were obtained by the similar image searching in the reference dataset. However, it's difficult for complicated indoor scene image to find a well matched

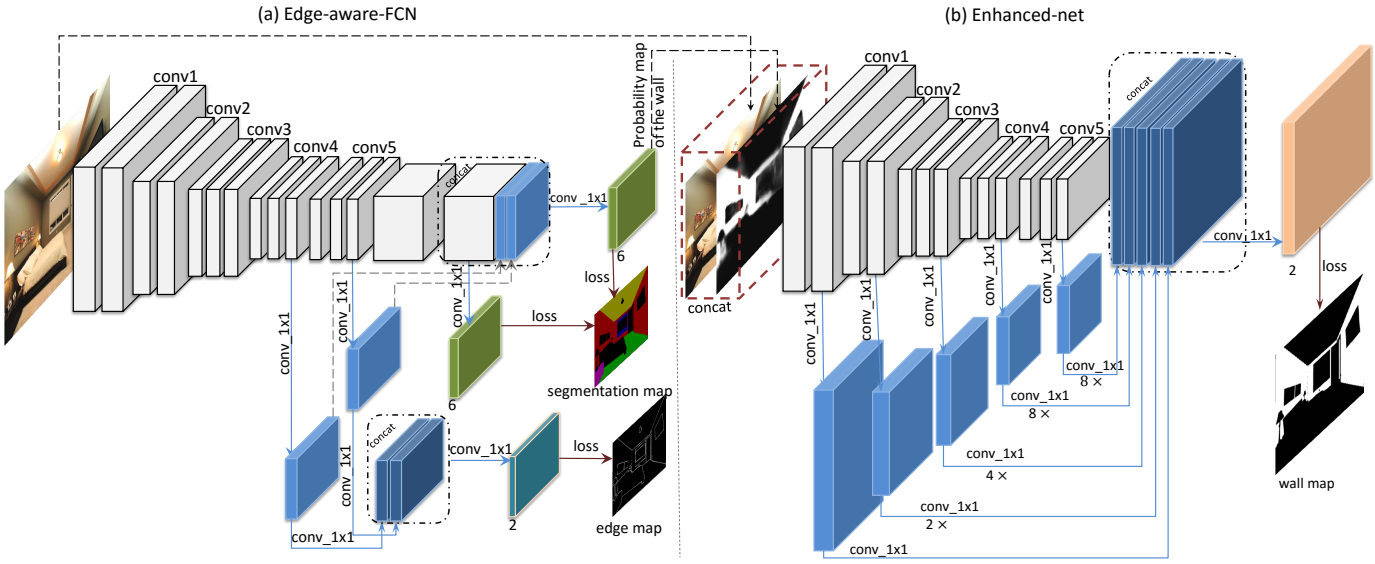


Fig. 2. Overview of the proposed Edge-aware-FCN and Enhanced-Net for magic-wall. The framework mainly consists of two components, Edge-aware-FCN and Enhanced-Net.Given an image, it's first fed into the Edge-aware-FCN for a coarse probability map. And then, the image with the corresponding probability map is fed to the next Enhanced-Net for more detailed prediction.

reference image. Hence, the target style is decided exclusively by users in our system. Recently, image style transfer [53], [54], [55], [56] has achieved great success as well with the deep learning techniques. Different from our semantic-aware stylization, the transformation was performed on the entire image. In addition, there existed several studies which aimed at assigning appearance to a 3D model, such as [57], [58], [59], [60], [61]. Specifically, Nguyen *et al.* [57] proposed an automatic system for assigning the material style from a guide image to a target 3D scene. Chen *et al.* [58] automatically generated material properties suggestions for all objects parts in the 3D scene. The 3D scene material assignment has a little parallelism with ours. However, the core problem of their works is to define the material and aesthetic rules that can be solved by combinatorial optimization.

III. Overview

Our system takes an indoor scene image as input and studies a semantics-driven approach to generate natural and vivid results with wall-color replaced. The system consists of two phases: a semantic content extraction phrase and a color replacement phrase.

In the first phase, the core idea is to accurately extract the semantic region of the wall for the input image. To tackle this issue, we propose an Edge-aware-FCN to parse the semantic contents and an Enhanced-Net to polish the details of extracted regions. The overall framework consisting of the two networks is shown in Fig. 2. For the Edge-aware-FCN, we learn edge detection and semantic prediction jointly and leverage the edge-aware features from the edge branch to refine the boundaries of semantic prediction. As shown in the left portion of Fig. 2, we predict edges by utilizing the front layers with low-level features and produce dense pixel-wise predictions from the last convolution layer. For Enhanced-Net, more front layers are adopted to further polish the obtained predictions produced by Edge-aware-FCN. Specifically, the probability map for the wall produced by Edge-aware-FCN is extracted at first. Then, the predicted probability map concatenated with RGB image is fed into the Enhanced-Net which could produce refined prediction by leveraging the features from the several front layers.

In the second phase, we focus on performing natural replacement for the obtained wall regions from the first phrase. We argue that the brightness information is crucial for the painted wall keeping in harmony with its background. For example, it's quite abrupt and inauthentic for an indoor scene that only the lamp is burning without any light and shadow on the wall. To strengthen the reality of the produced indoor scene image, we propose to employ a brightness reserved approach for replacement. We first transform RGB color space into HSV (hue, saturation, and value), in which the luminance channel is independent with others. Then, the entire replacement algorithm is implemented in HSV color space to reserve the luminance information of source image.

IV. Magic-wall Model

In this section, we first explain the proposed Edge-aware-FCN and Enhanced-Net more formally. And then, the details of color replacement with brightness reserved are introduced in the following.

Given the input training data set $\mathcal{D} = \{(X_n, Y_n, Z_n), n = 1, \ldots, N\}$, where X_n is the input image and $Z_n = \{z_j^n, j = 1, \ldots, |Z_n|\}, z_j^n \in \{0, 1\}$ is the corresponding binary edge map. $Y_n = \{y_j^n, j = 1, \ldots, |Y_n|\}, y_j^n \in \{0, 1, \ldots, C\}$ denotes the corresponding semantic map. Here, C is the number of categories of indoor scene, and 0 in denotes the background category. In the sequel, we omit the subscript n for notational convenience. Our target is to train an edge-aware semantic segmentation model $f(X; \theta)$ parameterized by θ and an

enhanced model $g(X;\phi)$ parameterized by ϕ. We minimize the following cross-entropy loss function to train our network.

$$\min_{\theta,\phi} \sum_{X\in\mathcal{D}} \underbrace{\mathcal{L}_{edge}(f(X;\theta)) + \mathcal{L}_{seg}(f(X;\theta))}_{\text{Edge-aware-FCN}} + \underbrace{\mathcal{L}_{wall}(g(X;\phi))}_{\text{Enhanced-Net}} \tag{1}$$

Here, the terms $\mathcal{L}_{edge}$ and $\mathcal{L}_{seg}$ are the loss functions of Edge-aware-FCN for edge prediction and pixel-wise semantic segmentation, respectively. $\mathcal{L}_{wall}$ is the loss function for Enhanced-Net.

A. Edge-aware-FCN

Edge Prediction The edge prediction branch is one of the components of the Edge-aware-FCN. We first connect convolutional layers with 1×1 kernel to last convolutional layer of the m^{th} convolutional group, which aims to extract features F_{e_m} for encoding useful information for both semantic segmentation and edge prediction. We then obtain multi-level 2-channel edge maps by performing 1×1 convolutional operations on the feature maps F_{e_m}. Those edge maps are further concatenated together and fused by a 1×1 convolution to generate 2-channels edge confidence map. We denote the final predicted edge map as P_e. During training, the loss function is computed over all pixels in a training image and edge map. Considering the heavily biased distribution of non-edge and edge pixels, we adopt a simply strategy class-balancing weight following [39]. We denote the non-edge and edge pixel sets as Z_0 and Z_1, respectively. The training loss used for this branch thus can be formulated as a class-balanced loss

$$\mathcal{L}_{edge} = \lambda \sum_{j\in Z_0} l(z_j, P_e(x_j)) + (1-\lambda) \sum_{j\in Z_1} l(z_j, P_e(x_j)), \tag{2}$$

where l is the cross-entropy loss defined on each pixel, z_j and $P_e(x_j)$ are ground truth label and predicted label for pixel x_j, respectively. λ is computed by the ratio of non-edge and edge pixels in the image, $\lambda = |Z_0|/|Z|$ and $1-\lambda = |Z_1|/|Z|$. Since our network aims at predicting semantic edge, the boundaries between different semantic regions in the mask, which does not correspond to traditional edge detection. Therefore, the high-level semantic features from end layers are required. Meanwhile, we need low-level features from front layers for edge detection. However, too much low-level features may detect the edge which is not corresponding to the semantic edge. Finally, we set $m = 4, 5$ experimentally in this paper since predicting from fourth and fifth convolutional group yields the best results for semantic segmentation in our experiments.

Semantic Segmentation In this part, the network generates semantic predictions from two branches. For the first non-edge-aware branch, we predict pixel-level semantic map P^{ne}_{seg} from the last convolutional layer *conv7* (denoted as F_{c_7}). In particular, we predict P^{ne}_{seg} by appending a 1×1 convolution layer with $(C+1)$ channels to *conv7*. For the edge-aware branch, we argue that the features from edge branch can learn different characteristics to complement the semantic features. Hence, we incorporate the semantic features F_{c_7} with edge-aware features F_{e_m} to produce better prediction P^{e}_{seg}. As shown in the left portion of Fig. 2, the features of intermediate layers from the edge branch are concatenated with that from *conv7* for making the dense pixel-wise prediction of semantic labels.

The two branches are optimized jointly to obtain more accurate prediction. We denote the following loss function for training

$$\mathcal{L}_{seg} = \sum\nolimits_j l(y_j, P^{ne}_{seg}(x_j)) + \sum\nolimits_j l(y_j, P^{e}_{seg}(x_j)), \tag{3}$$

where y_j is the ground-truth semantic label for pixel x_j. $P^{ne}_{seg}(x_j)$ and $P^{e}_{seg}(x_j)$ are the predicted label from non-edge-aware and edge-aware branch, respectively. The loss from non-edge-aware branch can encourage the feature maps from *conv7* to perceive high-level semantics of indoor scene. Since our target is to accurately locate pixels belonging to the wall for color replacement, we prefer the Edge-aware-FCN to pay more attentions on accurately segmenting 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, the weighted loss function is defined as

$$\mathcal{L}_{seg} = \sum\nolimits_j (l(y_j, P^{ne}_{seg}(x_j)) + l(y_j, P^{e}_{seg}(x_j))) \cdot \eta(y_j), \tag{4}$$

where $\eta(y_j)$ is the weight for label y_j. With the learned Edge-aware-FCN, the probability maps for the wall of the input image can be obtained, which are further fed into the following Enhanced-Net to polish the details.

B. Enhanced-Net

Even though the pixel-wise prediction generated from our Edge-aware-FCN is refined in the boundaries, it's still a little ambiguous in marginal regions that can't tally with the actual border. For instance, the border of the little photo frame hanging on the wall is hard to match with the actual borders completely. In this case, the fringe of the frame would be contaminated. [3] proposed improving localization performance by combining the responses at the final convolutional layer with a fully connected Conditional Random Field (CRF) [7]. However, the prediction is usually quite coarse in the blurred regions. Therefore, this post-processing method is not suitable for our specific application. To address the above issue, we propose an Enhanced-Net to polish the verge of the walls. Particularly, we leverage the predicted probability map from the previous stage as prior knowledge by concatenating with the RGB image to form a new input, and then feed it into our Enhanced-Net. In such a case, the Enhanced-Net focuses on polishing the more ambiguous details around the wall regions.

The input to the network is an image and the corresponding probability map(scaled between 0 and 255). Since we want the network to pay more attention to the details and localization, we employed the network structure similar to the edge detection in the first network. Hence, the shallow layers with more precise spatial information are leveraged to predicting together. As shown in Fig. 2, to capture more detailed features concurrently with high-level information, we first predict from the last feature maps of each convolutional group (from *conv1* and *conv5*) by conducting convolutional operations with 1×1 kernel size. Then the produced confidence

maps are concatenated together, which are further fused into 2-channel confidence map by 1×1 convolutional filters for wall prediction.

In this stage, we denote Y' as the corresponding pixel-wise wall region mask for input image I, in which the pixels belonging to wall regions are set as 1 and others are set as 0. Denoting the predicted wall map as P_{wall}, the loss used for optimizing the network can be defined as

$$\mathcal{L}_{wall} = \sum\nolimits_j l(y'_j, P_{wall}(x_j))) \tag{5}$$

All the parameter settings are detailed in Section 5. With the learned Enhanced-Net, the refined probability map for the wall of the testing image is extracted for the following substitution procedure.

C. Color/Texture Replacement

For wall color replacement, we use a technique by reserving the brightness of source image. Before replacing, we need to extract the wall mask for the input image. Usually, the predicted localization of the wall can't match the wall regions exactly. That is, there are always some pixels assigned improperly, especially in the blurred regions. Applying such hard assigned mask would result in inauthentic effect. Instead of a hard assigned mask, using probability maps can make the blurred regions have a smooth transition between wall and other semantic regions. Hence, we leverage the predicted probability map instead of the hard assigned mask to guide the entire substitution. For the color replacement, we generate a color image to have the same size as the input image, in which each pixel is valued by the new color. For the texture replacement, we obtain the final texture image by repeating it horizontally and vertically to have the same size with the input image.

Let I denotes the input image, and O denotes the output image. R represents the input color/texture map, which has the same size with I. Here, the confidence map P for the wall is exploited as the opacity for the new color/texture. The values of probability map P are scaled between 0 and 1, if a pixel in position x most likely belongs to the wall, $P(x)$ tends to be 1; otherwise, $P(x)$ tends to be 0. For each pixel x, we first use the following formulation to paint the wall regions:

$$O(x) = P(x) \cdot R(x) + (1 - P(x)) \cdot I(x) \tag{6}$$

Then, we use the following formulation in value channel to keep the brightness of source image:

$$O_V(x) = \omega \cdot P(x) \cdot R_V(x) + (1 - \omega) \cdot (1 - P(x)) \cdot O_V(x) \tag{7}$$

where ω is a hyper-parameter to control the extend of source illumination reserved. On one hand, we want to keep the source brightness as much as possible for authentic effect. On the other hand, large value will tarnish the reference color/texture. To balance the specified color/texture and original illumination, we set ω as 0.5 finally.

V. Experimental Results

A. Dataset and Settings

Dataset To benchmark our magic-wall system, we need collect a indoor scene dataset. To this end, we build a new dataset upon the publicly available ADE20K [1]. ADE20K contains a total 25K images of 150 classes with the semantic "wall" annotated, which just meets our requirements. To build our own dataset, we first removed all the images without "wall" semantic. And then, we select four semantic labels frequently appearing with the wall, 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. For edge branches, the ground truth of semantic edge is generated by annotating the boundaries between different semantic regions from the semantic label maps.

Implementation details Our network is implemented based on the publicly available Caffe library [62] and the open implementation of DeepLab [3]. As mentioned above, we take the VGG-16 as the backbone network, which is initialized with the pre-trained model provided by [3]. Following previous works, we train the network with standard stochastic gradient descent. Meanwhile, initial learning rate, momentum, and weight decay are set to 0.001, 0.9 and 0.0005, respectively. Besides, we adopt 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. The iteration number is set to 10K with a batch size of 4. Due to large cropsize can get good performance, the training images are cropped to 473×473, and the test images are cropped to 513×513. For comparison, we adopt Deeplab-LargeFOV as the baseline model. For Edge-aware-FCN, each intermediate convolutional layer of edge branches has 128 kernels of size 1×1. For Enhanced-Net, each intermediate convolutional layer has 32 kernels of size 1×1. To make the pre-trained model adapt 4-channel data, we initialize the one additional channel of the filters of the first convolutional layer with zeros.

During the test phase, given an input image, we first utilize the Edge-aware-FCN model to get the probability map for the wall, and then the probability map accompanying with the input image are fed into the Enhanced-Net for more accurate and detailed prediction. Finally, we extract the refined probability map to guide the entire substitution procedure.

Evaluation Metrics We adopt the intersection over union (IoU) for semantic segmentation evaluation in our experiments. Suppose C is the number of classes, $D_{i,j}$ is the number of pixels having ground-truth label i and whose prediction is j, we define $G_i = \sum_{j=1}^{C} D_{i,j}$, the total number of pixels labeled with i, and let $Pj = \sum_{i=1}^{C} D_{i,j}$ be the total number of pixels who were predicted as j. And then, IoU score can be computed with the following measure:

$$IoU = \frac{1}{C} \sum_{i=1}^{C} \frac{D_{i,j}}{G_i + P_i - D_{ii}} \tag{8}$$

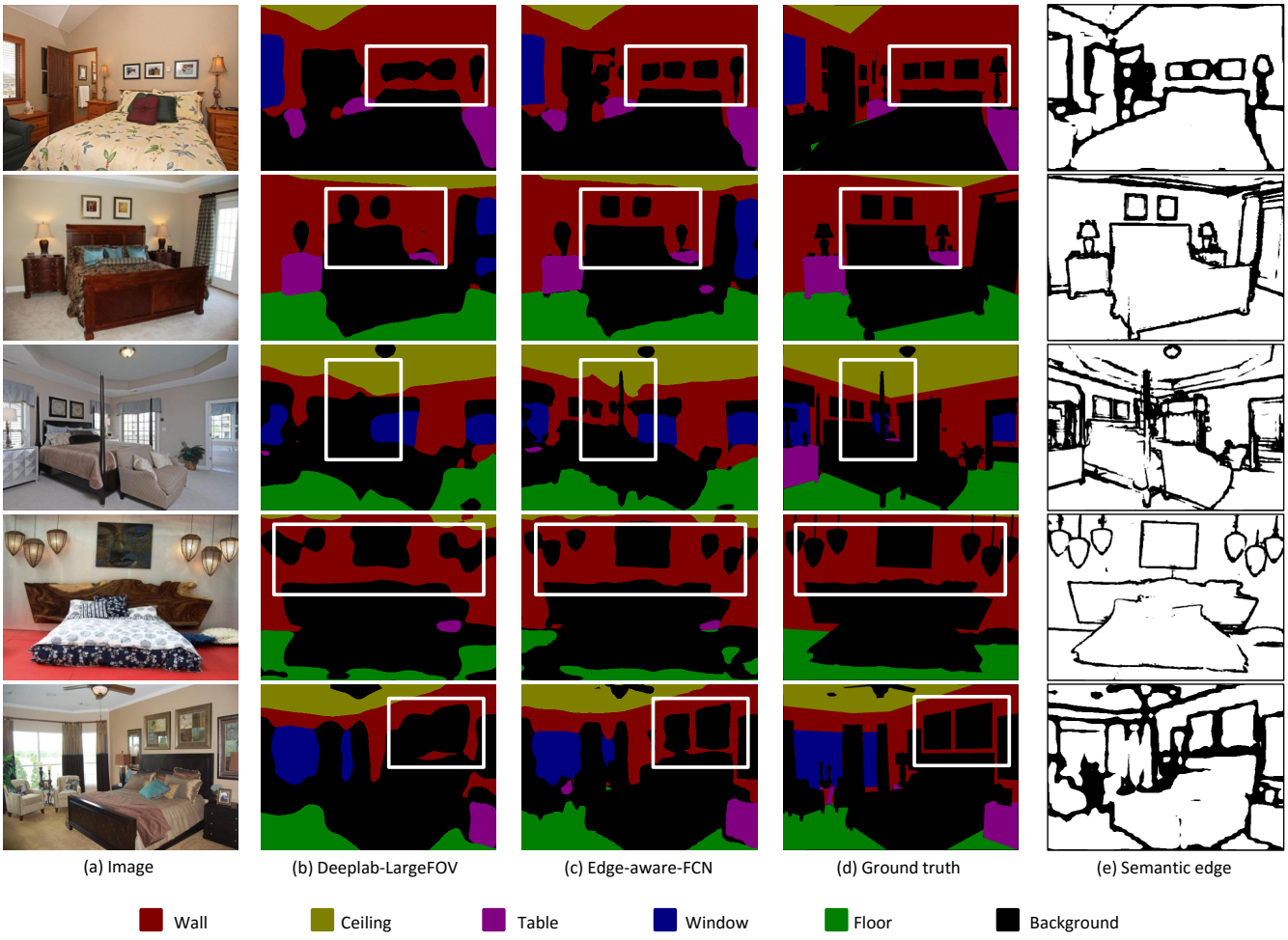


Fig. 3. Visualized semantic segmentation results. (a) The input image. (b) Scene parsing results of baseline network. (c) Scene parsing results of our Edge-aware-FCN. (d) The ground truth. (e) Edge detection results of our Edge-aware-FCN.

TABLE I
THE COMPARISON OF MIOU(%) FOR BASELINE AND OUR PROPOSED EDGE-AWARE-FCN BASE ON THE VGG-16 NETWORK.

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 (without non-edge-aware loss)	73.629	74.501	**69.929**	81.079	56.342	37.468	65.492
Edge-aware-FCN	**73.908**	**75.196**	69.518	**81.961**	**56.391**	**37.805**	**65.796**

Besides, to evaluate the replacement quality of our brightness reserved, we ask 23 participants to conduct the user study according to the verisimilitude of the composed images. In spired by [51], every participants are asked to rate each group for the harmony into five degrees: “much better”, “better”, “same”, “worse”, and “much worse”. At last, we count the percentage of every degree to report the performance of our proposed approach.

B. Ablation Analysis

In this section, we experiment with different design options to explore the effectiveness of each component of our method.

Classes To build the our dataset, we investigate the effect of the number of classes on the performance of the wall. We conduct experiments with the baseline model. The performances on the wall under different number of categories are reported in Tab. II. From the Tab. II, we can see that the number of categories has a very small impact on the results. Finally, we build the dataset with 6 classes since the performance on the wall is slightly better than the others.

Edge-aware-FCN We first compare our edge-aware-FCN with the baseline model to explore the effect of edge-aware features for semantic segmentation. We take the Deeplab-LargeFOV based on VGG-16 as the baseline model. And then, we implement the edge-aware-FCN shown in Fig. 2. To force the network focusing on the wall regions, we enlarge the loss

TABLE II
THE PERFORMANCE UNDER THE DIFFERENT NUMBER OF CLASSES.

Number of Classes	IoU(%)
2	72.432
6	**72.464**
20	72.217

weight $\eta(wall)$)(described in Eqn. 4) to 1.5. From the results reported in Tab. I, we can see that our outperforms the baseline model significantly. We get 65.796% on mIoU and 75.196% on the semantic of the wall, which are about 1.8% and 2.7% higher than the baseline model. The improvements well demonstrate that introducing the edge-aware feature is really beneficial for the semantic segmentation. Now we discuss the non-edge-aware loss. We evaluate by training with and without the non-edge-aware loss. From the results reported in Tab. I, we find that the performance on the wall decreases nearly 1%. Hence, we apply the non-edge-loss in the following experiments. Finally, we show several visualized results in Fig. 3 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 indicates that our model really captures the useful edge features and improves the semantic segmentation. In addition, we also display the semantic edge predicted by edge branches. Since we aim at predicting semantic edge to promote the semantic segmentation, the evaluation of the precision for the edge is not involved.

TABLE III
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

Edge Branches To better investigate the strength of the edge branches for our semantic segmentation, we use different edge architectures by employing various combinations of connections. Based on the baseline model, we vary the edge architectures by gradually introducing the front layers starting with *conv5*. The parameters λ in Eqn. 2 can be computed by the ratio of non-edge and edge pixels in the image. The performance of different combinations is listed in Tab. III. First, we can observe that all the combinations have improvements compared with the baseline performance. That's to say, the edge branches can capture discriminative features to promote the semantic segmentation. The combination of *conv4* and *conv5* yields the best performance. When we only utilize *conv5*, the mIoU is about 1% lower than the best performance. However, all also drop the performance about 1%. We analyze the reason is that our network aims at predicting semantic edge, the boundaries between different semantic regions in the mask, which does not correspond to traditional edge detection. The low-level edges are not included in this case, and the high-level semantic features are required to detect the semantic edges. Meanwhile, we also need the low-level features to capture details lost in high-level features. Hence, the combination of *conv4* and *conv5* boost the performance better.

TABLE IV
EFFECT OF OUR ENHANCED-NET. WE COMPARE THE RESULTS FROM EDGE-AWARE-FCN WITH/WITHOUT THE POST-PROCESSING OF ENHANCED-NET.

Network	IoU(%)
Edge-aware-FCN	75.196
Edge-aware-FCN + Enhanced-Net	**77.032**

Enhanced-Net The Enhanced-Net is based on the trimmed and dilated VGG-16 in which the first 14 convolutional layers are preserved. For training the Enhanced-Net, we extract the probability maps for all the images in the dataset with the trained Edge-aware-FCN model. The probability map and the image are incorporated into the input of the Enhanced-Net. Comparing the results before and after refining, the IOU on the wall is improved 1.8%. The reason is that the architecture of Enhanced-Net introduces more low-level features and focus on capturing the detailed information. In addition, aggregating the strong high-level probability map helps the network to focus on the ambiguous regions. The qualitative comparison of predicted results is visualized in Fig. 4. The visualized results show that our Enhanced-Net is effective in recovering the more detailed information. From the results before refining, we can see that the Edge-aware-FCN already fits the contours roughly. However, the boundaries along the wall are still quite vague. After employing the Enhanced-Net, the ambiguous regions, especially the edges of the wall, are identified legibly. To better show the effectiveness of our Enhanced-Net, a wall color replaced result with/without refining is shown in Fig. 5. Comparing the clock and branches in the red boxes, we can see that the regions with refining have sharper contours.

TABLE V
THE EFFECT OF DIFFERENT METHODS BASED ON VGG-16. **EDGE**: OUR EDGE-AWARE-FCN. **WEIGHT**: EMPLOYING WEIGHTED LOSS DURING TRAINING. **AUG**: DATA AUGMENTATION BY RANDOMLY RESCALING THE IMAGES. **COCO**: INITIALING WITH THE MODEL PRE-TRAINED ON MSCOCO. **ENHANCED**: REFINING WITH OUR ENHANCED-NET.

Edge	Weight	Aug	COCO	Enhanced	wall(%)	mIoU(%)
					72.464	63.967
✓					75.196	65.796
✓	✓				75.565	65.912
✓	✓	✓			76.022	67.441
✓	✓	✓	✓		76.412	67.915
✓	✓	✓	✓	✓	**77.549**	–

Other Tricks Besides, we adopt several other tricks during training following recent work of [3]: (1) Weighted loss: To further enhance the segmentation performance on the wall, we

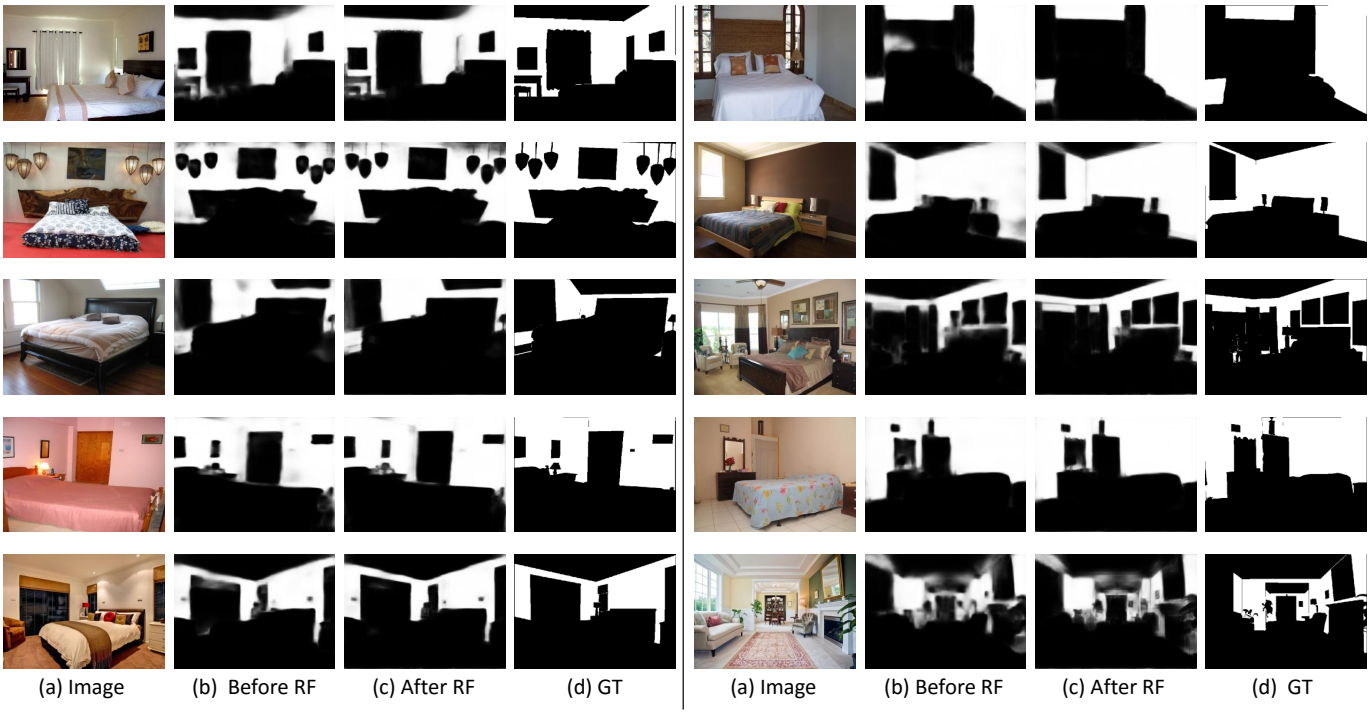


Fig. 4. Visual results of our Enhanced-Net. (a) Input image. (b) The probability map for the wall generated by our Edge-aware-FCN. (c) The probability map for the wall refined by our Enhanced-Net. (d) The ground truth.

Fig. 5. The visual effect of our Enhanced-Net. (a) Image with wall color replaced before refining. (b) Image with wall color replaced after refining. We can see the object boundaries in red boxes of the left image is much ambiguous than that of right.

want the network to pay more attention to the *wall* semantic. Therefore, weighted cross-entropy loss is introduced in the final network. (2) Data augmentation: Data augmentation has proven to be an effective approach to improve performance. Besides horizontal-flipping and random cropping, we also randomly scale the input images during each training iteration, and the scale is randomly sampled from 0.5, 0.75, 1, 1.25, 1.5. (3) pre-trained on MS-COCO [63], we further employ the initialized model which is pre-trained on MS-COCO dataset. In Tab. V, we show the evaluation of those tricks. From the Tab. V, we can see weighted loss results in 0.5 improvement on the wall. In the experiments, the weight on the wall is set as 1.5 empirically by the validation, and all others are set as 1. After applying the data augmentation, the mIoU is significantly improved 1.529% from 65.912% to 67.441%. The performance on the wall is slightly improved 0.5% as well. Additionally pre-trained on MS-COCO brings another near 0.5% improvement. Finally, with our Enhanced-Net and all the tricks mentioned above, the performance on the wall outperforms the baseline by 5.1%.

C. Evaluations on ResNet-101

As a better base network could gain significant improvement, ResNet-101 [9] is widely used in object detection and segmentation task. In order to further demonstrate the effectiveness of our proposed Edge-aware-FCN and Enhanced-Net, we also conduct experiments based on the ResNet-101. We adopt the Atrous Spatial Pyramid Pooling scheme for prediction like [3] as the baseline at this stage, which employs four parallel branches with different atrous rates in order to capture objects of different size.

For Edge-aware-FCN, we have demonstrated that the edge branches appended to the last two blocks yield the best performance. Hence, similar to the VGG-16, we attach the edge branches to the last convolutional layer in the last two blocks, respectively *res4f*, *res5c*. The feature maps of hidden layers in edge branches are further concatenated with *res5c* for jointly performing semantic segmentation. The results are shown as Tab. VI. From the Tab. VI, we can see that adopting ResNet-101 instead of VGG-16 improves the baseline mIoU from 65.796% to 68.647%. Even though ResNet delivers better segmentation results along object boundaries than employing VGG-16, our network still yields 1.763% (68.647% versus 70.41%) gap above the baseline with the proposed Edge-aware-FCN. The performance on the wall also achieves a 1.328% improvement above the baseline.

TABLE VI
THE COMPARISON OF MIOU(%) FOR BASELINE AND OUR PROPOSED EDGE-AWARE-FCN AND ENHANCED-NET BASE ON THE RESNET-101 NETWORK.

Method	Background	Wall	Floor	Ceiling	Window	Table	mIoU
Deeplab-ResNet	76.334	77.311	76.114	82.235	56.197	43.694	68.647
Edge-aware-FCN-ResNet	**77.659**	78.639	**77.721**	**83.233**	**58.312**	**46.899**	**70.410**
Edge-aware-FCN-ResNet + Enhanced-Net	–	**79.533**	–	–	–	–	–

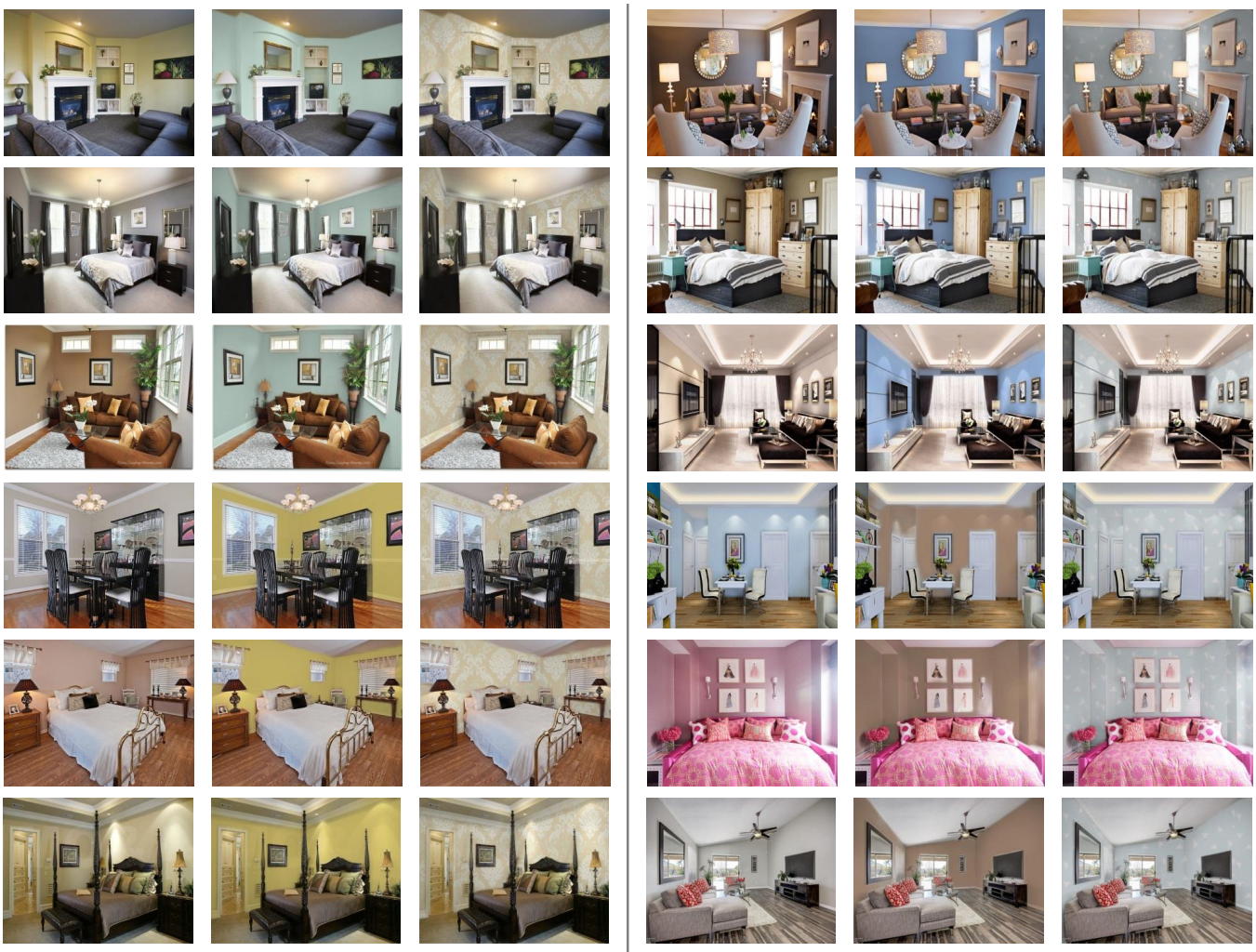


Fig. 6. Examples of indoor images with the color and texture of the wall replaced. (a) The indoor scene image under replaced. (b) The image with wall regions replaced by new color. (c) The image with wall regions replaced by new texture.

After applying the Edge-aware-FCN, the trained Enhanced-Net model is employed to polish the details. The performance of Enhanced-Net is provided in Tab. VI. With our Enhanced-Net, the performance attains to 79.533% outperforming near 0.9%. Since the ResNet-101 is better in localizing boundaries and our Edge-aware-FCN has enhanced the prediction marginal regions, the improvement is not significant compared with that based on VGG-16.

Finally, we extract the probability map from Edge-aware-FCN based on ResNet-101, and then employ the Enhanced-Net to polishing the details. After that, the refined probability maps are utilized to guide the following replacement procedure.

D. Replacement Results

To further validate the effectiveness of Magic-wall, several successful examples are shown together with color and texture replaced in Fig. 6. In addition, we conduct two groups of a user study to measure the quality of our approaches. In the experiment, we randomly select 70 images from the test set. Each time, a pair of images with different methods are sent to 23 participants to rate. For a fair comparison, note that the two images are shown randomly. Although such scores may vary across subjects, their statistics among 23 participants can still provide some useful cues to depict which method performs

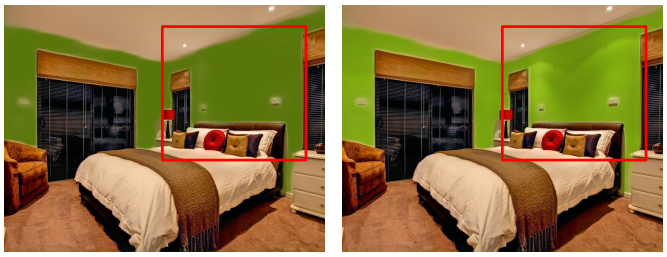


Fig. 7. Examples of replacement result with and without brightness reserved. From (a), we can see the illumination information especially around the lamp is lost. The wall with a new color is disharmonious with the entire indoor scene. From (b), with brightness reserved, the radiance of lights radiating to the wall is reserved.

Fig. 8. Visualization example of the conference version, Magic-wall-V1, and our approach. (a) The wall color replaced image by the Magic-wall-V1. (b) The wall color replaced image by our new approach.

better to some extent.

Brightness Reservation. The first group of user study aims at verifying the effectiveness of the brightness reserved. To balance the illumination between input image and reference color, the weight ω in Eqn. (7) is set as 0.5. As shown in Fig. 7(a), without brightness reserved, we can see that the glow of the lamp in Fig. 7(b) is reserved, which makes the replaced regions look more harmonious and verisimilar. The reason is that the Fig. 7(b) keeps the same illumination and shadow information with input image by reserving the original brightness, which is able to generate a more realistic result. Tab. VII reports the percentages of every degree comparing our approach with no brightness reserved and Magic-wall-V1. It can be found that the scores are mainly distributed in "better" compared with no brightness reserved, more than 56% is better or much better. It demonstrates that the brightness-reserved substitution is very critical to our magic-wall system.

Comparision with the Magic-wall-V1 The second group of user study compares the visualization results with our conference version Magic-wall-V1 [6]. Comparing with the Magic-wall-V1, we polish the details with the proposed Enhanced-net instead of the time-consuming CRF and Global Matting. In addition, we employ the probability map from the Enhanced-net for smoothly replacing rather than the hard assigned mask. As shown in VII, the scores of this work are concentrated on "better" comparing that with Magic-wall-V1. Our method is much better or better than Magic-wall-V1 in 1.59% and 45.09% cases. Since the above two methods bring smooth transition as shown in Fig. 8, the wall regions around the table lamps seem more naturally as the margin regions are smoothly replaced. The results demonstrate that our approach is able to generate more harmonious and realistic results.

TABLE VII
THE COMPARISON OF QUALITY FOR DIFFERENT METHODS. BR DENOTES BRIGHTNESS RESERVATION. MAGIC-WALL-V1 DENOTES THE RESULTS OF [6].

Method	much better	better	same	worse	much worse
Without BR	10.87%	45.9%	35.9%	5.47%	1.8%
Magic-wall-V1	1.59%	45.09%	27.76%	9.57%	1.68%

Besides, we report the comparison of the running time and performance for different methods in Tab. VIII. All the experiments are implemented on a Xeon E5-1650 v2 3.5GHz CPU and a GTX 1080 GPU. From Tab. VIII, we can see that the running time of Enhanced-Net is far faster than CRF. Moreover, the IoU on the wall of Enhanced-Net is higher than CRF 0.5%. Although the Edge-aware-FCN and Enhanced-Net need the training procedure, the training can be finished offline.

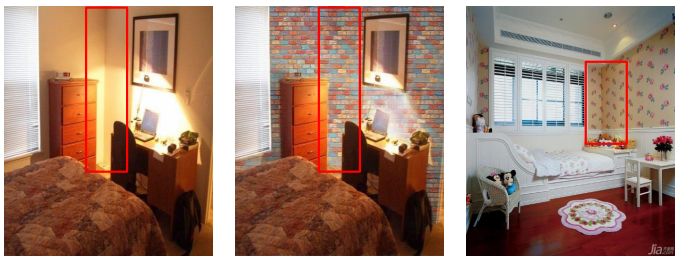
TABLE VIII
THE COMPARISON OF RUNNING TIME AND PERFORMANCE FOR DIFFERENT METHODS.

Method	training time (h)	test time (s)	IoU(%)
Edge-aware-FCN	2.8	0.044	75.196
CRF	-	0.656	76.534
Enhanced-Net	1.9	0.038	77.032

Limitation Although several excellent results replacing with texture have been given in Fig. 6, the texture replacement still has some problems. For instance, we show a failure case when replacing with texture in Fig. 9. In fact, there should have a warp when an image is taken from a 3D space. As shown in Fig. 9, we can see that the Fig. 9(b) lost the sense of space in the corner of the wall compared with Fig. 9(a), which makes the result a little unrealistic. The real texture on the wall should like the Fig. 9(c), in which the wallpaper has a deformation in the corner of the wall. On top of it all, the size of the pattern in the wallpaper should change from near to the distant like Fig. 9(c). To solve the existing issues, in the future work, we intend to leverage layout estimation [64], [65], [66] and homography matrix to warp the texture image for more naturalistic effect. Another problem is that we do not take the reflection of glass objects into account. As shown in Fig. 10, the color of the wall in the mirror is not replaced. Currently, our system still can't handle such a case.

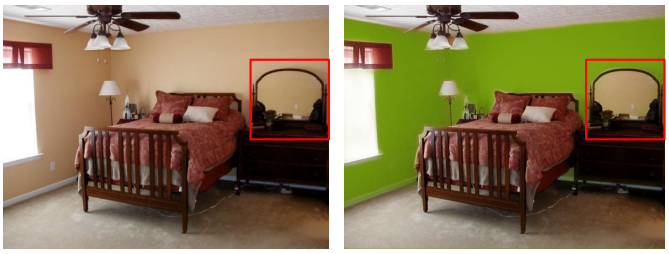
VI. CONCLUSION

In this paper, we have designed a Magic-wall system for automatically editing the wall color of an indoor scene image. To achieve this goal, we propose an edge-aware fully convolutional neural network and an enhanced network for



(a) Image (b) Texture replacement (c) Real world texture

Fig. 9. A failure example of texture replacement. (a) The input image. (b) Texture replacement with our proposed method. (c) A natural image of indoor scene with the real texture.



(a) Image (b) Color replaced

Fig. 10. A failure example of reflection. (a) The input image. (b) Color replaced result.

precisely localizing the wall region. Besides, we propose a color space conversion method for color replacing. The proposed Edge-aware-FCN shows significant improvement over the baseline deeplab-largeFOV. We also have demonstrated that the Enhanced-net is capable of polishing the details of the wall. Finally, we have shown in our experiments that our approach is able to compose more realistic results. Given that the current texture replacement will fail to generate a realistic result in some case, the layout of the indoor scene would help to overcome this issue. Our future work will consider making layout estimation for the indoor scene to perform natural texture replacement.

REFERENCES

[1] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[4] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *ECCV*. Springer, 2016, pp. 92–107.

[5] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *CVPR*, 2017.

[6] T. Liu, Y. Wei, Y. Zhao, S. Liu, and S. Wei, "Magic-wall: Visualizing room decoration," in *ACM MM*. ACM, 2017, pp. 429–437.

[7] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NIPS*, 2011, pp. 109–117.

[8] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, "A global sampling method for alpha matting," in *CVPR*, Jun. 2011, pp. 2049–2056.

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

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[13] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, Jul. 2017, pp. 2261–2269.

[14] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *CVPR*, 2017.

[15] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, and S. Yan, "Learning to segment with image-level annotations," *PR*, vol. 59, pp. 234–244, 2016.

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

[17] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *TPAMI*, vol. 38, no. 9, pp. 1901–1907, 2016.

[18] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *TGRS*, no. 99, pp. 1–13, 2018.

[19] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *T-ITS*, 2018.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *EC*. Springer, 2016, pp. 21–37.

[23] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *TPAMI*, 2018.

[24] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[25] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, Jul. 2017, pp. 5168–5177.

[26] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1520–1528. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.178

[27] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[28] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[30] S. Jgou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *CVPR*, Jul. 2017, pp. 1175–1183.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[32] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7268–7277.

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

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.

[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv:1802.02611*, 2018.

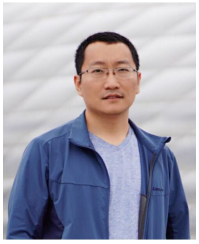
[36] Y. Ganin and V. Lempitsky, " nˆ 4-fields: Neural network nearest neighbor fields for image transforms," in *ACCV*. Springer, 2014, pp. 536–551.

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

[38] J. Hwang and T. Liu, "Pixel-wise deep learning for contour detection," *arXiv preprint arXiv:1504.01989*, 2015.

[39] S. Xie and Z. Tu, "Holistically-nested edge detection," *IJCV*, pp. 1–16, 2017.
[40] Y. Liu and M. S. Lew, "Learning relaxed deep supervision for better edge detection," in *CVPR*, 2016, pp. 231–240.
[41] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *CVPR*. IEEE, 2017, pp. 5872–5881.
[42] G. Bertasius, J. Shi, and L. Torresani, "High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision," in *ICCV*, 2015, pp. 504–512.
[43] I. Kokkinos, "Pushing the boundaries of boundary detection using deep learning," *ICLR*, 2015.
[44] L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *CVPR*, Jun. 2016, pp. 4545–4554.
[45] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," *TPAMI*, vol. 39, no. 1, pp. 115–127, 2017.
[46] J. M. Alvarez, A. M. Lopez, T. Gevers, and F. Lumbreras, "Combining priors, appearance, and context for road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1168–1178, 2014.
[47] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *T-ITS*, 2018.
[48] W. Song, L. Liu, X. Zhou, and C. Wang, "Road detection algorithm of integrating region and edge information," in *ICAIR*. ACM, 2016, p. 14.
[49] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, and M.-H. Yang, "Sky is not the limit: Semantic-aware sky replacement," *TOG*, vol. 35, no. 4, p. 149, 2016.
[50] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," in *CGF*, vol. 35, no. 2. Wiley Online Library, 2016, pp. 93–102.
[51] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, "Makeup like a superstar: Deep localized makeup transfer network," *arXiv preprint arXiv:1604.07102*, 2016.
[52] M. W. Tao, M. K. Johnson, and S. Paris, "Error-tolerant image compositing," in *ECCV*, 2010.
[53] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.
[54] ——, "Image style transfer using convolutional neural networks," in *CVPR*. IEEE, 2016, pp. 2414–2423.
[55] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images." in *ICML*, 2016, pp. 1349–1357.
[56] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016, pp. 694–711.
[57] C. H. Nguyen, T. Ritschel, K. Myszkowski, E. Eisemann, and H. Seidel, "3d material style transfer," in *CGF*, vol. 31, no. 2pt2, 2012, pp. 431–438.
[58] K. Chen, K. Xu, Y. Yu, T.-Y. Wang, and S.-M. Hu, "Magic decorator: automatic material suggestion for indoor digital scenes," *TOG*, vol. 34, no. 6, p. 232, 2015.
[59] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Opensurfaces: A richly annotated catalog of surface appearance," *TOG*, pp. 111:1–111:17, 2013.
[60] A. Jain, T. Thormählen, T. Ritschel, and H. P. Seidel, "Material memex: Automatic material suggestions for 3d objects," *TOG*, vol. 31, no. 5, 2012.
[61] M. G. Chajdas, S. Lefebvre, and M. Stamminger, "Assisted texture assignment," in *i3D*, 2010.
[62] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*. ACM, 2014, pp. 675–678.
[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.
[64] V. Hedau, D. Hoiem, and D. Forsyth, "Recovering the spatial layout of cluttered rooms," in *ICCV*, 2009.
[65] M. Arun and L. Svetlana, "Learning informative edge maps for indoor scene layout prediction," in *ICCV*, 2015, pp. 936–944.
[66] S. Dasgupta, K. Fang, K. Chen, and S. Savarese, "Delay: Robust spatial layout estimation for cluttered indoor scenes," in *CVPR*, 2016, pp. 616–624.

Ting Liu was born in Shannxi, China, in 1993. He received the B.S. degrees in computer science and technology in 2015, from the Beijing Jiaotong University of the Beijing, China. She is currently pursuing the Ph.D. degree with Institute of Information Science, Beijing Jiaotong University, Beijing, China. Her research interests include semantic segmentation and object detection.



Yunchao Wei received his Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2016, advised by Prof. Y. Zhao. He is currently a Postdoctoral Researcher at the University of Illinois at UrbanaChampaign. He received Excellent Doctoral Dissertation Awards of Chinese Institute of Electronics (CIE) in 2016, the Winner prize of the object detection task (1a) in ILSVRC 2014, the Runner-up prizes of all the video object detection tasks in ILSVRC 2017. He has published more than 30 papers in top-tier conferences/journals, with over 1000 citations in Google Scholar. His current research interest focuses on computer vision techniques for large-scale data analysis. Specifically, he has done work in weakly- and semi-supervised object recognition, multi-label image classification, image/video object detection, multi-modal analysis.

Yao Zhao (M'06−SM'12) received the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. In October 2015, he visited the Swiss Federal Institute of Technology, Lausanne, Switzerland (EPFL). From December 2017 to March 2018, he visited the University of Southern California. His current research interests include image/video coding, digital watermarking and forensics, video analysis and understanding and artificial intelligence. Dr. Zhao serves on the Editorial Boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor of Signal Processing: Image Communication. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a Fellow of the IET.

Si Liu (M12) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences. She was an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, and a Research Fellow with the Learning and Vision Research Group, Department of Electrical and Computer Engineering, National University of Singapore. She is currently an Associate Professor with the Beijing Key Lab Digital Media, School of Computer Science and Engineering, Beihang University. Her current research interests include object categorization, object detection, image parsing, and human pose estimation.

Shikui Wei received the B.E. degree from Hebei University and Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), China, in 2003 and 2010, respectively. From 2010 to 2011, he worked as a research fellow in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a full Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include computer vision, image/video analysis and retrieval, and machine learning. More information can be found at http://mic.bjtu.edu.cn.