

# SROBB: Targeted Perceptual Loss for Single Image Super-Resolution

Mohammad Saeed Rad<sup>1</sup>Behzad Bozorgtabar<sup>1</sup>Urs-Viktor Marti<sup>2</sup>Max Basler<sup>2</sup>Hazim Kemal Ekenel<sup>1,3</sup>Jean-Philippe Thiran<sup>1</sup><sup>1</sup>LTS5, EPFL, Switzerland<sup>2</sup>AI Lab, Swisscom AG, Switzerland<sup>3</sup>SiMiT Lab, ITU, Turkey

{saeed.rad, firstname.lastname}@epfl.ch

{firstname.lastname}@swisscom.com

## Abstract

By benefiting from perceptual losses, recent studies have improved significantly the performance of the super-resolution task, where a high-resolution image is resolved from its low-resolution counterpart. Although such objective functions generate near-photorealistic results, their capability is limited, since they estimate the reconstruction error for an entire image in the same way, without considering any semantic information. In this paper, we propose a novel method to benefit from perceptual loss in a more objective way. We optimize a deep network-based decoder with a targeted objective function that penalizes images at different semantic levels using the corresponding terms. In particular, the proposed method leverages our proposed OBB (Object, Background and Boundary) labels, generated from segmentation labels, to estimate a suitable perceptual loss for boundaries, while considering texture similarity for backgrounds. We show that our proposed approach results in more realistic textures and sharper edges, and outperforms other state-of-the-art algorithms in terms of both qualitative results on standard benchmarks and results of extensive user studies.

## 1. Introduction

Single image super-resolution (SISR) aims at solving the problem of recovering a high-resolution (HR) image from its low-resolution (LR) counterpart. SISR is a classic ill-posed problem that has been one of the most active research areas since the work of Tsai and Huang [33] in 1984. In recent years, this problem has been revolutionized by the significant advances in convolutional neural networks (CNNs) and has resulted in better reconstructions of high-resolution pictures than classical approaches [6, 5, 17]. More recently, another breakthrough has been made in SISR by employing perceptual loss functions for training feed-forward networks, instead of using per-pixel loss functions, e.g., mean squared error (MSE) [15, 27, 20]. It tackled the problem of blurred textures caused by optimization of MSE,

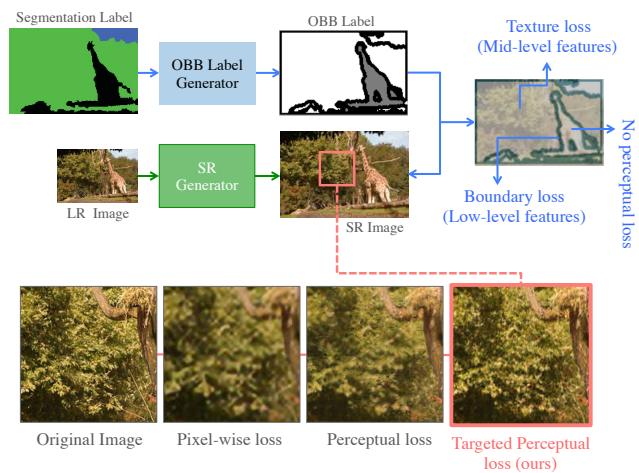


Figure 1. We propose a method for exploiting the segmentation labels during training to resolve a high resolution image at different semantic levels considering their characteristics; we optimize our SISR model by minimizing perceptual errors that correspond to edges only at object boundaries and the texture on the background area, respectively. Results from left to right: original image, super-resolved images using only pixel-wise loss function, pixel-wise loss + perceptual loss function and pixel-wise loss + targeted perceptual loss function (ours), respectively.

and alongside with adversarial loss [11], it resulted in near-photorealistic reconstruction in terms of perceived image quality.

[27] and [20] benefit from the idea of using perceptual similarity as a loss function; they optimize their models by comparing the ground-truth and the predicted super-resolved image (SR) in a deep feature domain by mapping both HR and SR images into a feature space using a pre-trained classification network. Although this similarity measure in feature space, namely the perceptual loss, has shown a great success in SISR, applying it as it is on a whole image, without considering the semantic information, limits its capability.

To better understand this limitation, let us have a brief

overview of the perceptual loss and see what a pre-trained classification network optimizes; considering a pre-trained CNN, in an early convolutional layer, each neuron has a receptive field with the size and shape of the inputs that affects its output. Small kernels, which are commonly used by state-of-the-art approaches, have also small receptive fields. As a result, they can only extract low-level spatial information. Intuitively, each neuron captures relations between nearby inputs considering their local spatial relations. These local relations are mostly presenting information about edges and blobs. As we proceed deeper in the network, the receptive field of each neuron with respect to earlier layers becomes larger. Therefore, deep layers start to learn features with global semantic meanings and abstract object information, and less fine-grained spatial details, while still using small kernels. This fact has also been shown by [40, 23], where they used some visualization techniques and investigated the internal working mechanism of the VGG network [29] by visualization of the information kept in each CNN layer.

Regarding the perceptual function, state-of-the-art approaches use different levels of features to restore the original image; this choice determines whether they focus on local information such as edges, mid-level features such as textures or high-level features corresponding to semantic information. In these works, perceptual loss has been calculated for an entire image in the same way, meaning that the same level of features has been used either on edges, foreground or on the image background. For example, minimizing the loss for details of the edges inside a random texture, such as the texture of a tree, would force the network to consider an unnecessary penalty and learn less informative features; the texture of a tree could still be realistic in the SR image without having close edges to the HR image. On the other hand, minimizing the loss by using mid-level features (more appropriate for the textures) around edges would not intuitively create sharper edges and would only introduce “noisy” losses.

To address the above issue, we propose a novel method to benefit from perceptual loss in a more objective way. Figure 1 shows an overview of our proposed approach. In particular, we use pixel-wise segmentation annotations to build our proposed OBB labels to be able to find targeted perceptual features that can be used to minimize appropriate losses to different image areas: e.g., edge loss for edges and textures’ loss for image textures during training. We show that our approach using targeted perceptual loss outperforms other state-of-the-art algorithms in terms of both qualitative results and user study experiments, and result in more realistic textures and sharper edges.

## 2. Related work

In this section, we review relevant CNN-based SISR approaches. This field has witnessed a variety of end-to-end deep network architectures: [17] formulated a recursive CNN and showed how deeper network architectures increase the performance of SISR. [20, 27, 45] used the concept of residual blocks [12] and skip-connections [13, 17] to facilitate the training of CNN-based decoders. [21] improved their models by expanding the model size. [36] removed batch normalization in conventional residual networks and used several skip connections to improve the results of seminal work of [20]. Laplacian pyramid structure [19] has been proposed to progressively reconstruct the sub-band residuals of high-resolution images. [31] proposed a densely connected network that uses a memory block consisting of a recursive unit and a gate unit, to explicitly mine persistent memory through an adaptive learning process. [44] proposed a channel attention mechanism to adaptively rescale channel-wise features by considering the inter-dependencies among channels. Besides supervised learning, other methods like unsupervised learning [41] and reinforcement learning [39] were also introduced to solve the SR problem.

Despite variant architectures proposed for the SISR task, the behavior of optimization-based methods is principally driven by the choice of the objective function. The objective functions used by these works mostly contain a loss term with the pixel-wise distance between the super-resolved and the ground-truth HR images. However, using this function alone leads to blurry and over-smoothed super-resolved images due to the pixel-wise average of all plausible solutions.

Perceptual-driven approaches added a remarkable improvement to image super-resolution in terms of the visual quality. Based on the idea of perceptual similarity [3], perceptual loss [15] is proposed to minimize the error in a feature space using specific layers of a pre-trained feature extractor, for example VGG [29]. A number of recent papers have used this optimization to generate images depending on high-level extracted features [9, 8, 38, 28, 34]. In a similar work, contextual loss [24] is proposed to generate images with natural image statistics, which focuses on the feature distribution rather than merely comparing the appearance. [20] proposed to use adversarial loss in addition to the perceptual loss to favor outputs residing on the manifold of natural images. The SR method in [27] develops a similar approach and further explores a patch-based texture loss. Although these works generate near-photorealistic results, they estimate the reconstruction error for an entire image in the same way, without benefiting from any semantic information that could improve the visual quality.

Many studies such as [7, 30, 32] also benefit from prior information for SISR. Most recently, [35] used an additional segmentation network to estimate probability maps as prior

knowledge and used them in the existing super-resolution networks. Their segmentation network is pre-trained on the COCO dataset [22] and then is fine-tuned on the ADE dataset [46]. Their approach recovers more realistic textures faithful to categorical priors; however, it requires a segmentation map at test-time. [26] addressed this issue by proposing a method based on multitask learning simultaneously for SR and semantic segmentation tasks.

In this work, we investigate a novel way to exploit semantic information within an image, yielding photo-realistic super-resolved images with fine-structures.

### 3. Methodology

Following recent approaches [20, 35, 25] for image and video super-resolution, we benefit from deep networks with residual blocks to build-up our decoder. As explained previously, in this paper, we focus on the definition of the objective function used to train our network; we introduce a loss function containing three terms: 1- Pixel-wise loss (MSE), 2- adversarial loss, and 3- our novel targeted perceptual loss function. The MSE and adversarial loss terms are defined as follows:

- **Pixel-wise loss** It is by far the most commonly used loss function in SR. It calculates the pixel-wise mean squared error (MSE) between the original image and the super-resolved image in the image domain [27, 5, 16]. The main drawback of using it as a stand-alone objective function is mostly resolving an over-smoothed reconstruction. The network trained with the MSE loss seeks to find pixel-wise averages of plausible solutions, which results in poor perceptual qualities and lack of high-frequency details in the edges and textures.
- **Adversarial loss** Inspired by [20], we formulate our SR model in an adversarial setting, which provides a feasible solution. In particular, we use an additional network (discriminator) that is alternatively trained to compete with our SR decoder. The generator (SR decoder) tries to generate fake images to fool the discriminator, while the discriminator aims at distinguishing the generated results from real HR images. This setting results in perceptually superior solutions to the ones obtained by minimizing pixel-wise MSE and classic perceptual losses. The discriminator used in this work is defined in more details in Section 3.3.

Our proposed targeted perceptual loss is described in the following subsection.

#### 3.1. Targeted perceptual loss

The state-of-the-art approaches such as [27] and [20] estimate perceptual similarity by comparing the ground-truth

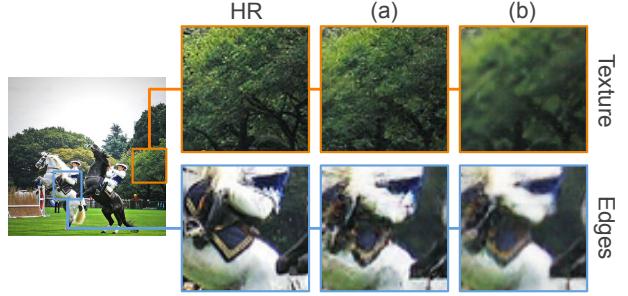


Figure 2. The effect of choosing different CNN layers to estimate the perceptual loss on different regions of an image, e.g., edges and textures: (a) using a deeper convolutional layer (mid-level features), ReLU 4-1 of VGG-16 [29] and, (b) using an early convolutional layer (low-level features), ReLU 1-2 of the VGG-16 network.

and the predicted super-resolved image in a deep feature domain by mapping both HR and SR images into a feature space using a pre-trained classification network, e.g., VGG [29]. The output of a specific convolutional layer is used as the feature map. These approaches usually minimize the  $l_2$  distance of the feature maps. In order to understand why minimizing this loss term in combination with adversarial and MSE losses is effective and results in more photorealistic images, we investigate the nature of the CNN layers used for the perceptual loss. Then, we propose a novel approach to take advantage of the perceptual similarity in a targeted manner and reconstruct more appealing edges and textures.

As explained previously, early layers of a CNN return low-level spatial information regarding local relations, such as information about edges and blobs. As we proceed towards deeper layers, we start to learn higher level features with more semantic meaning and abstract object information, and less fine-grained spatial details from an image. In this fashion, mid-level features are mostly representing textures and high-level features amount to the global semantic meaning. Figure 2 shows the difference between shallow and deep layers of a feature extractor, the VGG-16 in our case; two different layers, ReLU 1-2 and ReLU 4-1, are used to compute the perceptual loss and reconstruct an image. We compare each case on an edge and a texture region. In this figure, we can see using low-level features is more effective for reconstructing edges, while mid-level features resolve closer textures to the original image.

The targeted loss function tries to favor more realistic textures around areas, where the type of the textures seems to be important, e.g., a tree, while trying to resolve sharper edges around boundary area. To do so, we first define three types of regions in an image: 1- background, 2- boundaries, and 3- objects, then, we compute the targeted perceptual loss for each region using a different function.

- **Background** ( $\mathcal{G}_b$ ) We consider four classes as background: “sky”, “plant”, “ground” and “water”. We chose these categories because of their specific appearance; the overall texture in areas with these labels are more important than local spatial relations and edges. We compute mid-level CNN features to estimate the perceptual similarity between SR and HR images. Here, we use the ReLU 4-3 layer of the VGG-16 for this purpose.
- **Boundary** ( $\mathcal{G}_e$ ) All edges separating objects and the background are considered as boundaries. With some pre-processing (explained in more detail in Section 3.2), we broaden these edges to have a strip passing through all boundaries. We estimate the feature distance of an early CNN layer between SR and HR images, which focuses more on low-level spatial information, mainly edges and blobs. In particular, we minimize the perceptual loss at the ReLU 2-2 layer of the VGG-16.
- **Object** ( $\mathcal{G}_o$ ) Because of the huge variety of objects in the real world in terms of shapes and textures, it is challenging to decide whether it is more appropriate to use features from early or deeper layers for the perceptual loss function; for example, in an image of a zebra, sharper edges are more important than the overall texture. Having said that, forcing the network to estimate the precise edges in a tree could mislead the optimization procedure. Therefore, we do not consider any type of perceptual loss on areas defined as objects by weighting them to zero and rely on the MSE and adversarial losses. However, intuitively, resolving more realistic textures and sharper edges by the “background” and “boundary” perceptual loss terms would result in more appealing objects, as well.

To compute the perceptual loss for a specific image region, we make binary segmentation masks of the semantic classes (having a pixel value of 1 for the class of interest and 0 elsewhere). Each mask categorically represents a different region of an image and is element-wise multiplied by the HR image and the estimated super-resolved image SR, respectively. In other words, for a given category, the image is converted to a black image with only one visible area on it, before being passed through the CNN feature extractor. Masking an image in this way creates also new artificial boundaries between black regions and the visible class. As a consequence, extracted features contain information about the artificial edges which do not exist in a real image. As the same mask is applied on both HR and the reconstructed image, the feature distance between these artificial edges will be close to zero and it does not affect the total perceptual loss. We can conclude that all non-zero distances in feature

space between the masked HR and super-resolved image are corresponds to the contents of the visible area of that image: corresponds to edges by using a mask for boundaries ( $M_{OBB}^{boundaries}$ ) and corresponds to textures by using a mask for the background ( $M_{OBB}^{background}$ ).

The overall targeted perceptual loss function is given as:

$$\begin{aligned}\mathcal{L}_{perc.} = & \alpha \cdot \mathcal{G}_e(I^{SR} \circ M_{OBB}^{boundary}, I^{HR} \circ M_{OBB}^{boundary}) \\ & + \beta \cdot \mathcal{G}_b(I^{SR} \circ M_{OBB}^{background}, I^{HR} \circ M_{OBB}^{background}) \\ & + \gamma \cdot \mathcal{G}_o\end{aligned}\quad (1)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are the corresponding weights of the loss terms used for the boundary, background, and object, respectively.  $\mathcal{G}_e(\cdot)$ ,  $\mathcal{G}_b(\cdot)$  and  $\mathcal{G}_o(\cdot)$  are the functions to calculate feature space distances between any two given images for the boundaries, background, and objects, respectively. In this equation,  $\circ$  denotes element-wise multiplication. As discussed earlier, we do not consider any perceptual loss for objects areas, therefore, we set  $\gamma$  directly to zero. The value of other weights are discussed in detail in Section 4.1.

In the following subsection, we describe how to build a label indicating objects, the background, and boundaries for the training images. This labeling approach helps us to use specific masks for each class of interest ( $M_{OBB}^{object}$ ,  $M_{OBB}^{background}$  and  $M_{OBB}^{boundary}$ ) and to guide our proposed perceptual losses to focus on area of interest within the image.

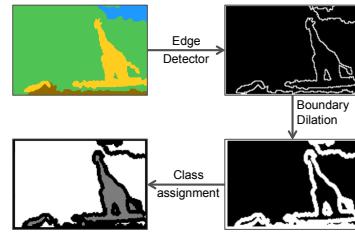


Figure 3. Constructing an OBB label. We assign each area to one of the “Object”, “Background” or “Boundary” classes based on their initial pixel-wise labels.

### 3.2. OBB: Object, background and boundary label

In order to make full use of the perceptual loss-based image super-resolution, we enforce semantic details (where objects, the background, and boundaries appear on the image) via our proposed targeted loss function. In addition, existing annotations for the segmentation task, e.g., [4] only provide spatial information about objects and the background, and they do not use classes representing the edge areas, namely boundaries in this paper. Therefore, inspired by [26], we propose our labeling approach (Figure 3) to provide a better spatial control of the semantic information for the images.

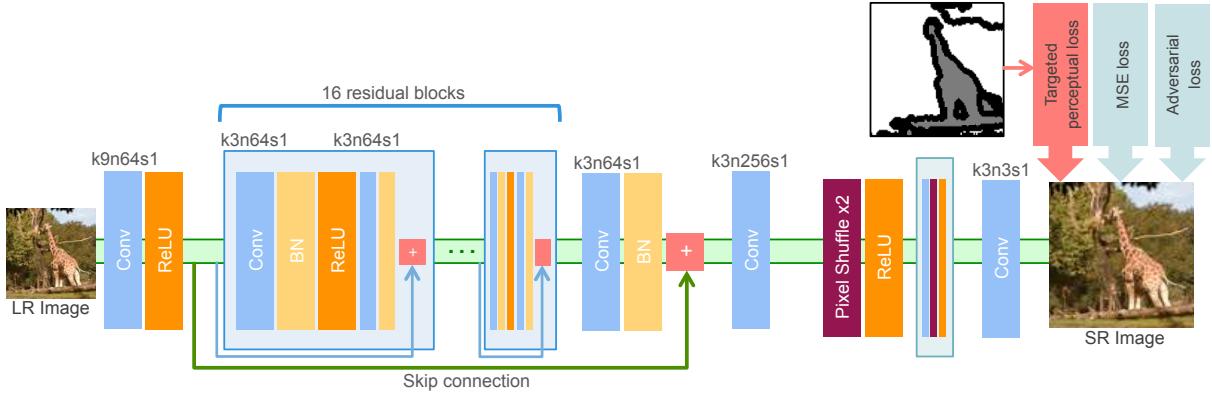


Figure 4. Schematic diagram of the SR decoder. We train the SR decoder using the targeted perceptual loss alongside with MSE and adversarial losses. In this schema,  $k$ ,  $n$  and  $s$  correspond to kernel size, number of feature maps and stride size, respectively.

To create such labels (OBB label), first, we calculate the derivative of the segmentation label in the color-space to estimate the edges between object classes in the segmentation label as well as the edges between objects and background of the image. In order to have a thicker strip around all edges separating different classes, we compute the dilation with a disk of size  $d_1$ . We label the resulted area as “boundary” class, which covers boundaries between different classes inside an image. In particular, we consider “sky”, “plant”, “ground”, and “water” classes from the segmentation labels as the “Background”. All remaining object classes are considered as the “object” class.

### 3.3. Architecture

For a fair comparison with the SRGAN method [20] and performing an ablation study of the proposed targeted perceptual loss, we use the same SR decoder as the SRGAN. The generator network is a feed-forward CNN. The input image  $I^{LR}$  is passed through a convolution block followed by a ReLU activation layer. The output is subsequently passed through 16 residual blocks with skip connections. Each block has two convolutional layers with  $3 \times 3$  filters and 64 channels feature maps, each one followed by a batch normalization and ReLU activation. The output of the final residual block is concatenated with the features of the first convolutional layer and is then passed through two upsampling blocks, where each one doubles the size of the feature map. Finally, the result is filtered by a last convolution layer to get the super-resolved image  $I^{SR}$ . In this paper, we use a scale factor of four; depending on the desired scaling factor, the number of upsampling blocks could be modified. An overview of the architecture is shown in Figure 4.

The discriminator network consists of multiple convolutional layers with an increasing number of channels of the feature maps by a factor of 2, from 64 to 512. We use Leaky-ReLU and strided convolutions to reduce the image dimension while doubling the number of features. The re-

sulting 512 feature maps are passed through two dense layers. Finally, the discriminator network classifies the image as real or fake by the final sigmoid activation function.

## 4. Experimental Results

In this section, first, we describe the training parameters and dataset in details, then we evaluate our proposed method in terms of qualitative, quantitative, and running costs analysis.

### 4.1. Dataset and parameters

To create OBB labels, we use a random set of 50K images from the COCO-Stuff dataset [4], which contains semantic labels of 91 classes for the segmentation task. In this paper, we considered landscapes with one or more of the “Sky”, “Plant”, “Ground”, and “Water” classes. We group these classes into one “Background” class. We use our proposed technique in Section 3.2 to convert pixel-wise segmentation annotations to OBB labels. In order to obtain LR images, we use the MATLAB imresize function with the bicubic kernel and the anti-aliasing filter. All experiments were performed with a downsampling factor of four.

The training process was done in two steps; first, the SR decoder was pre-trained for 25 epochs with only pixel-wise mean squared error as the loss function. Then the proposed targeted perceptual loss function, as well as the adversarial loss were added and the training continued for 55 more epochs. The weights of each term in the new targeted perceptual loss,  $\alpha$  and  $\beta$ , were set to  $2 \times 10^{-6}$  and  $1.5 \times 10^{-6}$ , respectively. The weights of adversarial and MSE loss function, as in [20], were set to 1.0 and  $1 \times 10^{-3}$ , respectively. We set  $d_1$ , the diameter of the disk used to generate OBB labels, to 2.0. The Adam optimizer [18] was used during both steps. The learning rate was set to  $1 \times 10^{-3}$  and then decayed by a factor of 10 every 20 epochs. We also alternately optimized the discriminator with similar parameters to those proposed by [20].

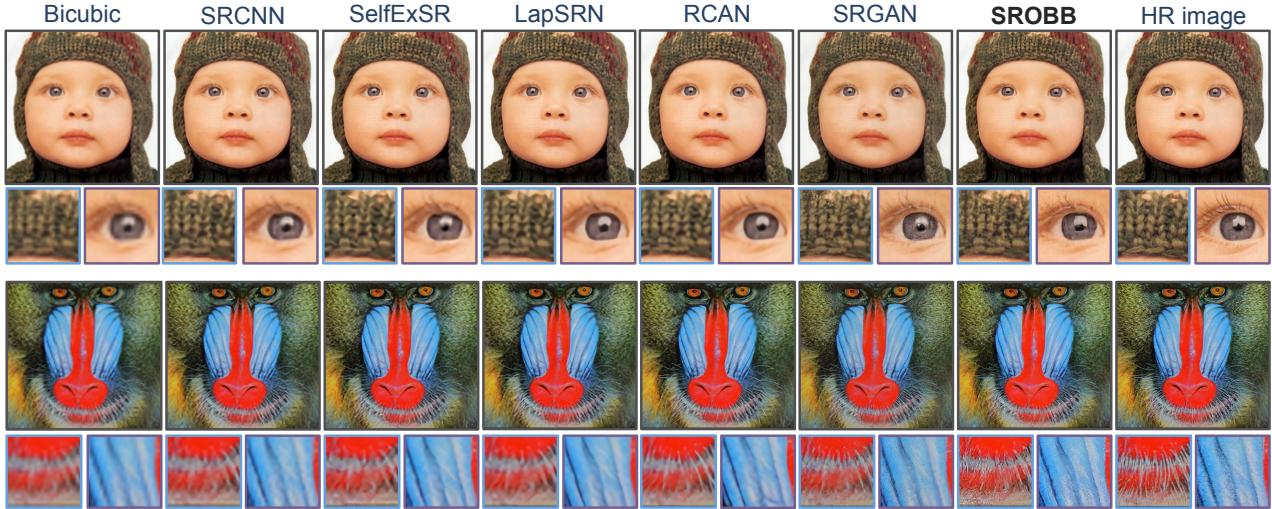


Figure 5. Sample results on the “baby” (top) and “baboon” (bottom) images from Set5 [1] and Set14 [2] datasets, respectively. From left to right: bicubic, SRCNN [5], SelfExSR [14], LapSRN [19], RCAN [44], SRGAN [20] and SROBB (ours), HR image, respectively.

## 4.2. Qualitative Results

### 4.2.1 Results on Set5 and Set14

Our approach focuses on optimizing the decoder with perceptual loss terms targeting boundaries and background by exploiting segmentation labels. Although, we do not apply the perceptual losses specifically on objects regions, our experiment shows that the trained model generalized in a way that it reconstructs more realistic objects compared to other approaches. We evaluate the quality of object reconstruction by performing qualitative experiments on two widely used benchmark datasets: Set5 [1] and Set14 [42], where unlike our training set, in most of the images, outdoor background scenes are not present. Figure 5 compares the results of our SR model on the “baby” and “baboon” images and the recent state-of-the-art methods including: bicubic, SRCNN [5], SelfExSR [14], LapSRN [19], RCAN [44] and SRGAN [20]. In the “baboon” image, we could generate more photo-realistic images with sharper edges compared to other methods while having competitive results for the “baby” image with SRGAN. Their results were obtained by using their online supplementary materials<sup>1 2 3</sup>. More qualitative results of Set5 and Set14 images are provided in the supplementary material.

### 4.2.2 Results on the COCO-Stuff dataset

We randomly chose a set of test images from the COCO-Stuff dataset [4]. In order to have a fair comparison, we re-trained the SFT-GAN[35], ESRGAN [36] and SRGAN

[20] methods on the same dataset with the same parameters as ours. For the EnhanceNet and RCAN, we used their pre-trained models by [27] and [44], respectively. The MATLAB imresize function with a bicubic kernel is used to produce bicubic images. As illustrated in Figure 6, our method generates more realistic and natural textures by benefiting from our proposed targeted perceptual loss. Although ESRGAN produces very competitive results, it seems that their method is biased towards over-sharpened edges, which sometime leads to an unrealistic reconstruction and dissimilar to ground-truth.

## 4.3. Quantitative Results

### 4.3.1 SSIM, PSNR and LPIPS

As it is shown in [20, 27, 35, 2], distortion metrics such as the Structural Similarity Index (SSIM) [37] or the Peak Signal to Noise Ratio (PSNR) used as quantitative measurements, are not directly correlated to the perceptual quality; they demonstrate that GAN-based super-resolved images could have higher errors in terms of the PSNR and SSIM metrics, but still generate more appealing images.

In addition, we used the perceptual similarity distance between the ground-truth and super-resolved images. The Learned Perceptual Image Patch Similarity (LPIPS) metric [43] is a recently introduced as a reference-based image quality assessment metric, which seeks to estimate the perceptual similarity between two images. This metric uses linearly calibrated off-the-shelf deep classification networks trained on the very large Berkeley-Adobe Perceptual Patch Similarity (BAPPS) dataset [43], including human perceptual judgments. However, as [10] also emphasizes, LPIPS has similar trend as distortion-based metrics, e.g., SSIM, and would not necessarily imply photorealistic images.

<sup>1</sup><https://github.com/jbhuang0604/SelfExSR>

<sup>2</sup><https://github.com/phoenix104104/LapSRN>

<sup>3</sup><https://twitter.app.box.com/s/lcue6vlrd011jktdkxmfvk7vtjhetog>

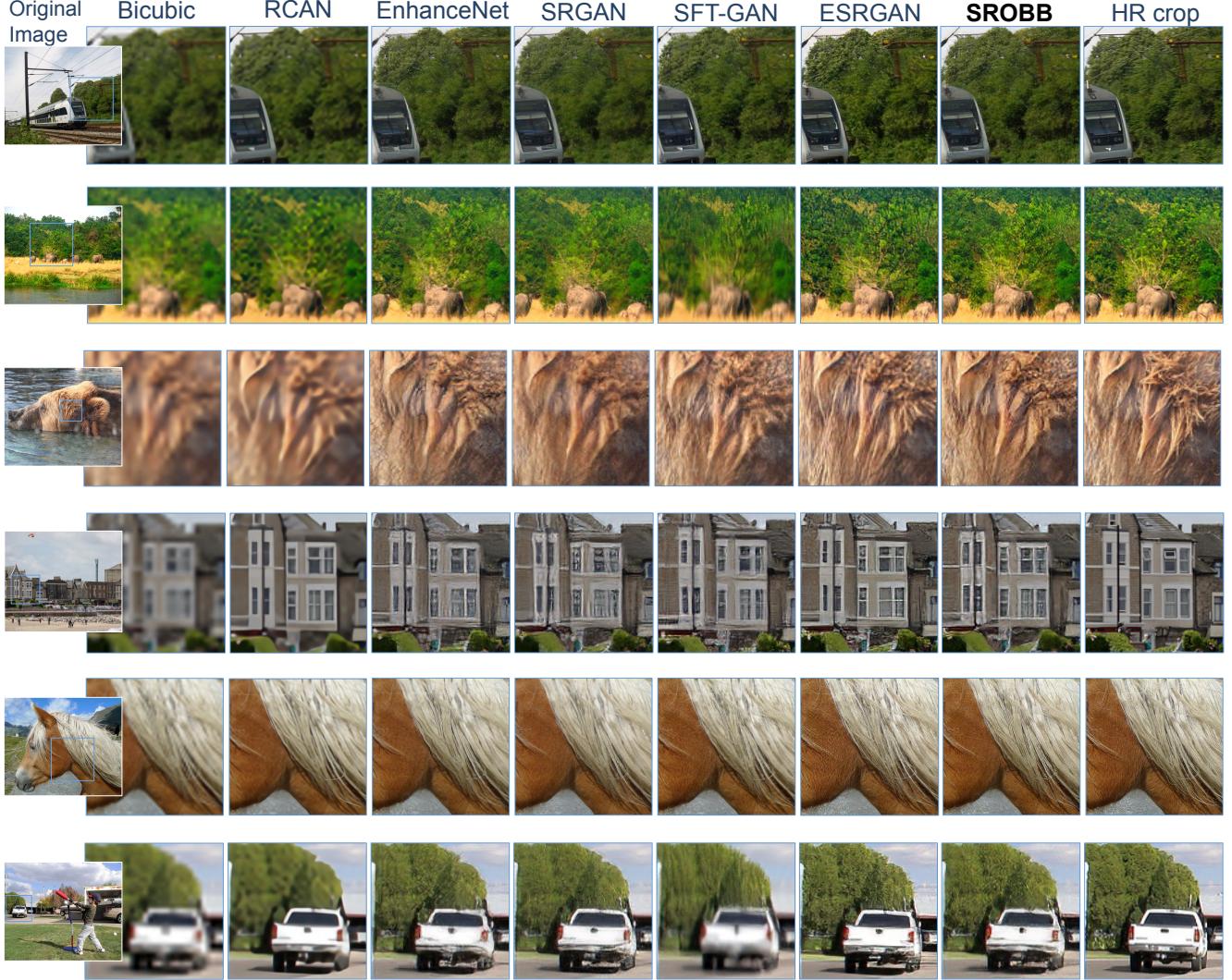


Figure 6. Qualitative results on a subset of the COCO-Stuff dataset [4] images. Cropped regions are zoomed in with a factor of 2 to 5 to have a better comparison. Results from left to right: bicubic, RCAN [44], EnhanceNet [27], SRGAN [20], SFT-GAN [35], ESRGAN [36], SROBB (ours) and a high resolution image. Zoom in for the best view.

Image	Metric	Bicubic	LapSRN	SRGAN	SROBB
baby	SSIM	0.936	<b>0.951</b>	0.899	0.905
	PSNR	30.419	<b>32.019</b>	28.413	28.869
	LPIPS	0.305	0.237	0.112	<b>0.104</b>
baboon	SSIM	0.645	<b>0.677</b>	0.615	0.607
	PSNR	20.277	<b>20.622</b>	19.147	18.660
	LPIPS	0.632	0.537	<b>0.220</b>	0.245

Table 1. Comparison of bicubic interpolation, LapSRN [19], SRGAN [20] and SROBB (ours) for the “baby” and “baboon” images from Set5 and Set14 test sets. Best measures (SSIM, PSNR [dB], LPIPS) are highlighted in bold. The visual comparison is shown in Figure 5.

Table 1 shows the SSIM, PSNR, and LPIPS values estimated between super-resolved images of the “baby” and “baboon” and their HR counterparts, using bicubic interpolation, LapSRN [19], SRGAN [20], and our method, respectively. Considering this table and the visual comparison of these images in Figure 5, we can infer that these metrics would not reflect superior reconstruction quality. Therefore, in the following section, we focus on the user study as the quantitative evaluation.

#### 4.3.2 User study

We performed a user study to compare the reconstruction quality of different approaches to see which images are more appealing to users. Five methods were used in the

study: 1- RCAN [44], 2- SRGAN [20], 3- SFT-GAN [35], 4- ESRGAN [36] and 5- SROBBB (ours). During the experiment, high-resolution images as well as their five reconstructed counterparts obtained by the mentioned approaches were shown to each user. Users were requested to vote for more appealing images with respect to the ground-truth image. In order to avoid random guesses in case of similar qualities, a choice as “Cannot decide” was also designed. Since SFT-GAN uses a segmentation network trained on outdoor categories, for a fair comparison with [35], we also used 35 images from COCO-Stuff [4], dedicated to outdoor scenes. All images were presented in a randomized fashion to each person. In order to maximize the number of participants, we created our online assessment tool for this purpose. In total, 46 persons participated in the survey. Figure 7 illustrates that the images reconstructed by our approach are more appealing to the users by a large margin. In terms of number of votes per method, reconstructions by the SROBB got 617 votes, while ESRGAN, SFT-GAN, SRGAN and RCAN methods got 436, 223, 201 and 33 votes, respectively. In addition, the “Cannot decide” choice provided in the survey was chosen 100 times. In terms of the best images by majority of votes, among 35 images, SROBB was a dominant choice in 15 images. These results confirm that our approach reconstructs visually more convincing images compared to mentioned methods for the users. Moreover, unlike SFT-GAN, the proposed approach do not require a segmentation map during the test time, while it takes advantage of semantic information and produces competitive results.

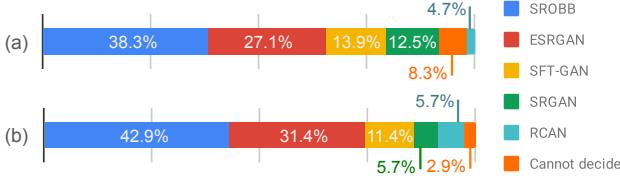


Figure 7. The results of the user study, comparing SROBB (ours) with RCAN [44], SRGAN [20], ESRGAN [36] and SFT-GAN [35] methods. Our method produces visual results that are the preferred choice for the users by a large margin in terms of: (a) percentage of votes, (b) percentage of winning images by majority of votes.

### 4.3.3 Ablation study

To better investigate the effectiveness of the proposed targeted perceptual loss, we performed a second user study with similar conditions and procedure to the one in the previous section. Specifically, we study the effect of our proposed targeted perceptual loss; we train our decoder with three different objective functions: 1- pixel-wise MSE only; 2- pixel-wise loss and standard perceptual loss similar to

[20]; and 3- Pixel-wise loss and our proposed targeted perceptual loss (SROBB). The adversarial loss term is also used for both 2 and 3. In total, 51 persons participated in our ablation study survey. Figure 8 shows that users are more convinced when the targeted perceptual loss is used instead of the commonly used perceptual loss. It got 1212 votes, while objective functions 1 and 2 got 49 and 417 votes, respectively. In addition, the “Cannot decide” choice was chosen 107 times. In terms of the best images by majority of votes, among 35 images, third objective function was a dominant choice in 30, while 1 and 2 won only in 5 images. Images reconstructed only by the pixel-wise loss had minority number of votes, however, they got considerable number of votes for images in which the “sky” was the main class. This can be explained by the over-smoothed nature of the clouds, which suits distortion-based metrics.

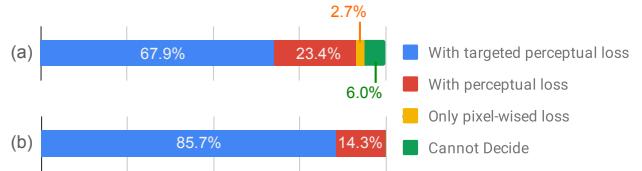


Figure 8. The results of the ablation study showing the effect of the targeted perceptual loss; more convincing results have been obtained by a large margin, in terms of: (a) percentage of votes, (b) percentage of winning images by majority of the votes.

### 4.4. Inference time

Unlike existing approaches for content-aware SR, our method does not require any semantic information at the input. Therefore, no additional computation is needed at the test time. We reach an inference time of 31.2 frame per second, with a standard XGA output resolution ( $1024 \times 768$  pixels) on a single GeForce GTX 1080 Ti.

## 5. Conclusion

In this paper, we introduced a novel targeted perceptual loss function for the CNN-based single image super-resolution. The proposed objective function penalizes different regions of an image with the relevant loss terms, meaning that using edges’ loss for the edges and textures’ loss for textures during the training process. In addition, we introduce our OBB labels, created from pixel-wise segmentation label, to provide a better spatial control of the semantic information for the images. This allows our targeted perceptual loss to focus on the semantic regions of an image. Experimental results verify that training with proposed targeted perceptual loss yields perceptually more pleasing results, and outperforms the state-of-the-art SR methods.

## References

- [1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *British Machine Vision Conference (BMVC)*, Guildford, Surrey, United Kingdom, Sept. 2012.
- [2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. 2018 PIRM challenge on perceptual image super-resolution. *CoRR*, abs/1809.07517, 2018.
- [3] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. *arXiv preprint arXiv:1511.05666*, 2015.
- [4] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2016.
- [5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2014.
- [6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 184–199, Cham, 2014. Springer International Publishing.
- [7] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, July 2011.
- [8] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.
- [9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. *CoRR*, abs/1505.07376, 2015.
- [10] Muhammad Waleed Gondal, Bernhard Schölkopf, and Michael Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution. In *ECCV Workshops*, 2018.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *ECCV*, 2016.
- [14] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2015.
- [17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2015.
- [18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.
- [19] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843, 2017.
- [20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2016.
- [21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017.
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [23] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233255, May 2016.
- [24] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Learning to maintain natural image statistics. *arXiv preprint arXiv:1803.04626*, 2018.
- [25] Eduardo Pérez-Pellitero, Mehdi S. M. Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Photorealistic video super resolution. In *Workshop and Challenge on Perceptual Image Restoration and Manipulation (PIRM) at the 15th European Conference on Computer Vision (ECCV)*, 2018.
- [26] Mohammad Saeed Rad, Behzad Bozorgtabar, Claudio Musat, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Benefiting from multitask learning to improve single image super-resolution. *accepted at Neurocomputing (Special Issue on Deep Learning for Image Super-Resolution)*, 2019.
- [27] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through

- automated texture synthesis. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4501–4510, 2016.
- [28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.
- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [30] J. Sun, J. Zhu, and M. F. Tappen. Context-constrained hallucination for image super-resolution. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 231–238, June 2010.
- [31] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4539–4547, 2017.
- [32] Radu Timofte, Vincent De Smet, and Luc Van Gool. Semantic super-resolution: When and where it is useful? *Computer Vision and Image Understanding*, September 2015.
- [33] R. Tsai and T. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, 1:317339, 1984.
- [34] Subeesh Vasu, Thekke Madam Nimisha, and A. N. Rajagopalan. Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In *ECCV Workshops*, 2018.
- [35] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018.
- [36] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018.
- [37] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.
- [38] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.
- [39] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2443–2452, 2018.
- [40] Wei Yu, Kuiyuan Yang, Yalong Bai, Hongxun Yao, and Yong Rui. Visualizing and comparing convolutional neural networks, 2014.
- [41] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 814–81409, 2018.
- [42] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proceedings of the 7th International Conference on Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer-Verlag.
- [43] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [44] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.
- [45] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [46] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4. IEEE, 2017.