Sky is Not the Limit: Semantic-Aware Sky Replacement

Yi-Hsuan Tsai[1*] Xiaohui Shen[2] Zhe Lin[2] Kalyan Sunkavalli[2] Ming-Hsuan Yang[1]
[1]University of California, Merced [2]Adobe Research

Input photograph Automatic sky replacement results with different styles

Figure 1: *Given an input photograph, the proposed algorithm automatically generates a set of images with stylized skies. The proposed algorithm exploits visual semantics for sky editing, in which scene parsing is first performed on the input image, and such semantic information is utilized for the subsequent steps including sky segmentation, search and replacement.* *Photo credits: daveynin, Dilexa, John Williams and Michele Landi.*

Abstract

Skies are common backgrounds in photos but are often less interesting due to the time of photographing. Professional photographers correct this by using sophisticated tools with painstaking efforts that are beyond the command of ordinary users. In this work, we propose an automatic background replacement algorithm that can generate realistic, artifact-free images with a diverse styles of skies. The key idea of our algorithm is to utilize visual semantics to guide the entire process including sky segmentation, search and replacement. First we train a deep convolutional neural network for semantic scene parsing, which is used as visual prior to segment sky regions in a coarse-to-fine manner. Second, in order to find proper skies for replacement, we propose a data-driven sky search scheme based on semantic layout of the input image. Finally, to re-compose the stylized sky with the original foreground naturally, an appearance transfer method is developed to match statistics locally and semantically. We show that the proposed algorithm can automatically generate a set of visually pleasing results. In addition, we demonstrate the effectiveness of the proposed algorithm with extensive user studies.

Keywords: sky segmentation, sky replacement, compositing, appearance transfer, semantic search

Concepts: •**Computing methodologies** → **Image processing;** *Computational photography;*

1 Introduction

Skies are one of the most common backgrounds in photos. However, we have no control over the weather or lighting conditions at the moment of photography. As a result, numerous interesting and valuable photos have uninteresting or poorly exposed sky regions. Professional photographers fix this problem using sophisticated tools by manually delineating the sky regions precisely, testing different skies for compatibility, and finally adjusting the foreground to match the new composited sky. This requires time and expertise that is beyond the abilities of novice users. In this paper, we propose a fully automatic sky replacement tool that can take an input image and generate a diverse set of realistically edited photos with interesting skies and different styles (see Figure 1). This can expedite the editing process for professionals, and allow casual users with minimal expertise to explore interesting results.

To achieve this goal, we address three challenging questions in this work. Can we accurately segment the sky region from the image? How do we find interesting sky images that are compatible with the input image? Finally, can we match the appearance of the input and new sky images to create realistic composites? In this paper, we show that a deep semantic understanding of images is critical to all these tasks. For example, sky appearance varies widely among images, and without an understanding of scene layout, it can be indistinguishable from non-sky regions (e.g., reflections in water). Similarly, it is important that we search photos whose semantic layout and content match the input image. This ensures that the perspective of the composited sky is consistent with the input image. It also allows us to adjust the foreground appearance after sky replacement and improve the realism of the result. For instance, when adjusting the color of water under a new sky, it is better to transfer appearance from a water region appearing with that sky than from an arbitrary region with other content.

In this work, we propose a semantic-aware approach for sky editing, in which visual semantics are extracted from an input image, and instilled into the subsequent steps: sky segmentation, search and replacement. The overall framework of our approach is illustrated in Figure 2. We first learn a deep Fully Convolutional Neural Network (FCN) [Long et al. 2015] to parse an input image and generate a dense pixel-wise prediction of semantic labels such as sky, tree, building, mountain and water. By understanding the global layout of the scene, the proposed algorithm can robustly localize the sky regions in spite of different colors, shapes, sizes and attributes. The proposed online classifier is then trained to model the sky appearance and used to generate a refined sky segmentation mask with clean boundaries across the sky and non-sky regions. Experiments show that the generated sky segmentation results by the proposed

*e-mail:ytsai2@ucmerced.edu

SIGGRAPH '16 Technical Paper, July 24-28, 2016, Anaheim, CA,
ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925942

algorithm achieve high intersection-over-union (IOU) ratio against the ground truth masks, thereby making it effective at identifying background regions without manual refinement.

To select interesting and stylized sky exemplars for replacement, we construct a database with 415 high aesthetic quality images covering a wide range of sky appearance and scene categories. We use the FCN trained for scene parsing to construct feature vectors that represent the semantic content and layout of those exemplars, which enables us to retrieve images that are semantically similar to an input photo yet have diverse styles of sky regions. Once the sky images are selected, the new sky regions are automatically aligned and composited into the input photo. The color, saturation and luminance of the non-sky region are then adaptively adjusted to ensure that the composite photos are visually realistic. Since the content of the reference images are similar to the input image, and the semantic regions have already been segmented on both the input and reference images, we leverage this information and propose a novel semantic-aware approach to transfer the appearance of the reference exemplars to the input and create results with stylized sky backgrounds.

We demonstrate that the proposed algorithm is able to render photorealistic and pleasing images through extensive user studies. In particular, we show that our approach performs favorably against existing methods for scene compositing in terms of realism, diversity, and interestingness. The contributions of this work for automatic semantic-aware sky replacement are summarized as follows:

1. We propose a fully automatic sky replacement algorithm that is driven by scene semantics.
2. We develop an accurate coarse-to-fine sky segmentation method using a deep neural network and online classifier learning (Section 4).
3. We show a semantic search scheme to determine a set of high-quality images with diverse sky appearance and similar semantic content for replacement (Section 5).
4. We present a semantic-aware appearance transfer approach for replacing sky backgrounds to render images that are aesthetically pleasing and photorealistic (Section 6).

2 Related Work

The goal of this work is to create realistic composites where the original sky region of an input image is replaced by more interesting and stylized backgrounds. This task entails a combination of high-quality segmentation, semantic matching, and appearance transfer. In this section, we discuss existing methods closely related to these modules within the context of rendering images with stylized sky backgrounds.

Appearance Matching for Compositing. Creating realistic composites requires a good match between both the content and the appearance of the images being merged. When the images being merged are already specified, existing techniques use color and tone matching to ensure that the generated results have consistent appearance. Color and tone matching can be carried out by transferring global statistics [Reinhard et al. 2001; Pitié and Kokaram 2007] or using correspondences between local regions [Tai et al. 2005; HaCohen et al. 2011]. Gradient domain schemes have been developed to address inconsistencies on boundaries [Pérez et al. 2003; Tao et al. 2013]. In addition, the problems with textural inconsistency between images can be alleviated by matching multi-scale statistics [Sunkavalli et al. 2010] or patch-level adjustments [Darabi et al. 2012].

While these methods directly match the appearance of images being composited, another class of techniques learns the transformations from external datasets. Xue et al. [2012] learn color and tone transformations such that the appearance of the composited foreground regions is adjusted properly with respect to the background. Color and tone transformations have also been exploited to hallucinate changes in time of day [Shih et al. 2013] and other high-level transient attributes [Laffont et al. 2014]. On the other hand, data-driven techniques have been proposed to restore degraded photographs [Dale et al. 2009] and improve the realism of computer generated images [Johnson et al. 2011]. Lalonde and Efros [2007] learn color statistics from a set of natural images to predict the realism of photos, and use them to adjust foregrounds to improve the chromatic compatibility. Recently, Zhu et al. [2015] extend this work with a convolutional neural network (CNN) to learn a model for predicting and improving the realism of composites. Moreover, a CNN based method [Yan et al. 2016] that utilizes semantic features is developed to learn appearance adjustment, in which pairs of input and output styles are required for training the CNN.

Semantic Search for Compositing. Appearance matching methods perform well when the content of the images being considered are consistent, and numerous search techniques have been developed to determine compatible images. Hays et al. [2007] present an image completion method by searching a large database for images with compatible layout measured by the GIST descriptor [Torralba et al. 2006]) and appearance. To composite images, Lalonde et al. [2007] search for objects that are consistent with the input photograph in terms of camera orientation, lighting, resolution, etc. While the above-mentioned techniques consider generic objects, Bitouk et al. [2008] propose a method that specifically replaces faces with compatible pose, appearance and lighting. Note that all these approaches rely on hand-crafted features to find compatible images. In this work, we use a deep neural network to extract semantic features. In addition, we leverage semantic information as a core component to search for sky exemplars, as well as the subsequent segmentation and appearance matching that are essential to create high-quality composites.

Closest in scope to our work is the work of Tao et al. [2009] who present a system to search for skies with specified attributes. However, the quality of sky replacement is measured by a simple geometric metric to score compatibility, and global color transfer is used to match image appearance. In contrast, we exploit scene semantics which allow us to design refined segmentation, search, and local appearance transfer algorithms to generate more realistic composite results.

Semantic Segmentation. Sky detection or segmentation has been addressed for sky image search [Tao et al. 2009], sky model estimation [Lalonde et al. 2010] and photo enhancement [Kaufman et al. 2012], in which hand-crafted features including color, position and texture are used together with traditional rule-based or learning-based classifiers such as SVM. On the other hand, sky region can also be detected and segmented by performing semantic scene parsing on the image, which helps better understand the overall image content and layout [Hoiem et al. 2007]. Representative methods of scene parsing include exemplar based label transfer [Liu et al. 2011] and superpixel matching [Tighe and Lazebnik 2013].

Recently, Convolutional Neural Network (CNN) based methods [Long et al. 2015; Chen et al. 2015; Zheng et al. 2015] for semantic segmentation have drawn attention due to their significantly improved accuracy and scalability. In particular, fully-connected Conditional Random Field (dense CRF) utilized in [Chen et al. 2015] refines the results generated by CNN. In addition, Zheng et al. [2015] make the dense CRF a trainable module to enable end-to-end training. Instead of using dense CRF, we propose to

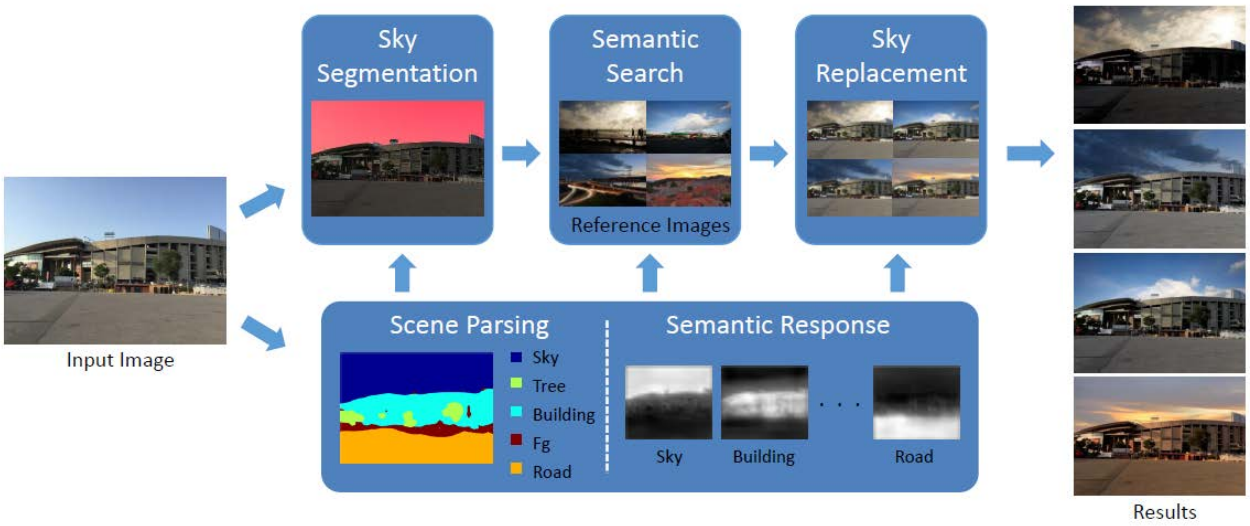


Figure 2: *Overview of the proposed algorithm. Given an input image, we first utilize the FCN to obtain scene parsing results and semantic response for each category. A coarse-to-fine strategy is adopted to segment sky regions (illustrated as the red mask). To find reference images for sky replacement, we develop a method to search images with similar semantic layout. After re-composing images with the found skies, we transfer visual semantics to match foreground statistics between the input image and the reference image. Finally, a set of composite images with different stylized skies are generated automatically.* *Photo credits: Scott Cohen, PokoPoko, Yikuen Tsai, Alex and Danny Molyneux.*

learn an online discriminative model initialized by the CNN outputs for sky segmentation refinement, which achieves better accuracy as demonstrated in our experiments. It is in spirit similar to the object-specific segmentation method via Markov random fields [Kumar et al. 2005] and online discriminative models [Rother et al. 2004; Liu and Yu 2012] used in interactive segmentation. Our approach differs from these methods in that we formulate the task in a coarse-to-fine manner, in which we utilize a CNN to first localize the sky regions, and our online classifiers are specifically designed based on the properties of sky appearance.

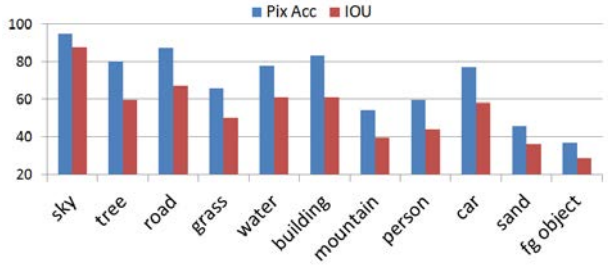


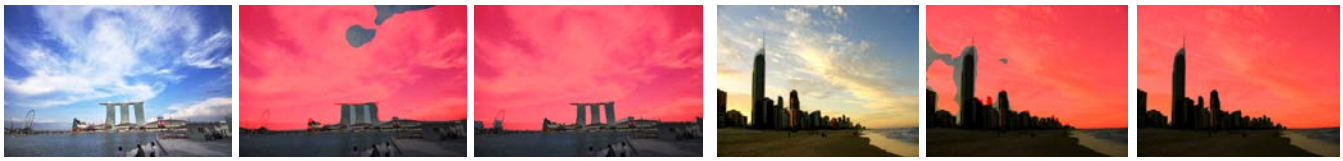
Figure 3: *Pixel accuracy and intersection over union ratio of the scene parsing results for each category on the LMSun dataset.*

3 Algorithmic Overview

Given an input image I, we aim to automatically generate a set of results with the same foreground as in I but different sky backgrounds. To achieve this, we decompose the task into three subtasks: (1) sky segmentation for separating the sky region I^{sky} and the foreground region I^{fg} in image I; (2) sky search for retrieving a set of reference images matching the input image semantics; (3) sky replacement for replacing the original sky region I^{sky} with a new sky R^{sky}, given a reference image R. Meanwhile, the color statistics of the foreground region I^{fg} are automatically adjusted to ensure the composed images are visually consistent and realistic.

The main idea of our work is to exploit visual semantics to help improve the image quality in all three tasks. Toward this, we train a Fully Convolutional Neural Network (FCN) [Long et al. 2015] for scene parsing, which is a state-of-the-art end-to-end model for semantic segmentation (see Figure 2 for an example). To learn an accurate FCN for scene parsing, we randomly select 15000 outdoor images form the LMSun dataset [Tighe and Lazebnik 2013] as training samples. As the labels in the LMSun dataset include many small objects or labels that do not appear often in our daily photos, we manually choose the common labels that cover most scene categories, and merge other object labels to one category as the foreground object. Figure 3 shows the list of 11 pre-defined labels for the image editing task considered in this paper. In this work, the semantic deep neural network is trained in a way similar to [Long et al. 2015]. We quantitatively evaluate on 1045 randomly selected images form the LMSun dataset, and compute the pixel accuracy and intersection-over-union (IOU) ratio. Figure 3 shows that semantic segmentation with the defined labels can be accurately computed by the trained neural network. This model can robustly localize arbitrary skies and effectively distinguish the true sky from those most confusing regions such as reflections in the water and mountains in the hazy background. Based on the coarse scene parsing results, we delineate more accurate sky regions with clear segmentation boundaries by online refinement (Section 4).

The FCN output provides visual presentations describing the scene layout, which facilitates the subsequent sky search and replacement steps. In particular, we normalize the semantic response maps generated by the FCN forward propagation over all the categories to obtain a probability map $F_i = \{f_i^1, f_i^2, \ldots, f_i^n\}$, where f_i^n indicates the likelihood of pixel i belonging to category n. We construct a semantic layout descriptor based on the probability map for the input image as well as for all the images in the sky database, which is used to retrieve a set of images from the database that have similar content and layout to the input image. The skies in those images are therefore likely compatible with the input foreground region, and are proper for replacement. Moreover, since we are only using semantic information for retrieval instead of traditional appearance features, we can find images with different sky appearance to ensure



(a) Input image (b) FCN result (c) Fine segmentation (d) Input image (e) FCN result (f) Fine segmentation

Figure 4: *Sample sky segmentation results. Given an input image, the FCN generates results that localize the sky well but contain inaccurate boundaries and noisy segments. The proposed online model refines segmentations that are complete and accurate, especially around the boundaries (best viewed in color with enlarged images).* *Photo credits: d.i. and motiqua.*

the diversity of our results. The algorithmic details of the proposed semantic search scheme are described in Section 5.

Once a reference image is selected, its sky region will be used to replace the original one. As the appearance of the sky is dramatically changed, the color statistics of the foreground need to be adjusted accordingly to make the image visually realistic. However, global transfer techniques are susceptible to differences in the foreground and can create non-realistic results. In contrast, we have semantic segmentation of the scene that we use to drive a local transfer technique that produces more realistic results. The results show that our transfer method can generate realistic and visually pleasing results compared to other methods (Section 6). Figure 2 shows the main steps of our method for replacing sky backgrounds based on image semantics.

4 Sky Segmentation

Based on the scene parsing results by the trained FCN, we first introduce an accurate sky segmentation algorithm to facilitate sky replacement and reduce visual artifacts. As discussed in Section 3, the scene parsing model trained by FCN can robustly localize the sky regions with various appearance and scene layout. In particular, it can achieve 94% pixel accuracy on our evaluation set. Nevertheless, since the image resolution of the FCN output is low, the resulting segmentation masks are coarse with missing details around the boundaries, which cannot be directly used for replacement (see Figure 4).

Sky Segment Refinement via Online Models. In order to generate accurate sky segmentation masks, we use the FCN results to bootstrap online classifiers that learn image-specific color and texture models. We formulate a two-class CRF problem for refinement by considering neighboring pixels x_i and x_j with the energy $E(X)$,

$$E(X) = \lambda_1 \sum_i U_c(x_i) + \lambda_2 \sum_i U_t(x_i) + \lambda_3 \sum_i U_f(x_i) + \lambda_4 \sum_{(i,j)\in\mathcal{E}} V(x_i, x_j), \quad (1)$$

where U_c and U_t are color and texture unary potentials for the cost to be the sky or non-sky labels, which are obtained from the learned online classifier, and U_f is a location term that accounts for the FCN output. In addition, V is the pairwise potential for smoothness in a set $\mathcal{E}$ of adjacent pixels, and $\lambda's$ are the weights for each term. Here we use equal weights for three unary terms ($\lambda_1 = \lambda_2 = \lambda_3 = 1$), and a higher weight ($\lambda_4 = 100$) for the pairwise term to ensure the boundary smoothness.

To learn online sky/non-sky classifiers and model unary potentials, we first use the sky response map generated by the FCN output layer as sky/non-sky priors. More specifically, if a pixel has higher response (e.g., larger than a threshold) on the sky label, we use this pixel as the positive training sample to learn the sky classifier, and vice versa. We consider two cues for learning online models. First, we use the Gaussian Mixture Models (GMMs) on RGB channels to model the appearance of sky regions, and obtain U_c by computing the negative logarithm of GMM probability outputs.

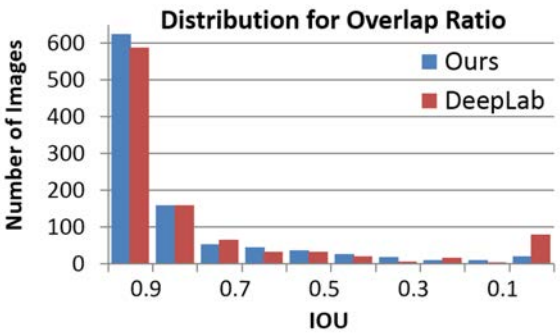


Figure 5: *Distribution of images in terms of IOU ratio compared to the DeepLab method [Chen et al. 2015]. Most sky segmentation results by the proposed algorithm have over* 90% *IOU.*

However, when the color of sky regions are similar to foreground ones, the CRF model with only the chromatic cues may generate noisy segmentation results. Hence, we also model the texture of sky regions by computing histogram of gradients on superpixels, and learn a Support Vector Machine (SVM) classifier. Similarly, U_t is obtained by calculating the negative logarithm of the SVM outputs that are converted to probabilities through a sigmoid mapping function. Note that the texture model is based on superpixels, and we assign the energy to each pixel according to its parent region and then perform pixel-wise energy minimization.

In addition, we use the sky response map from the FCN to guide the sky location for each pixel x_i, where $U_f(x_i)$ is equal to the negative logarithm of response f_i^{sky}. For the pairwise term V, we use the magnitude of gradient between two adjacent pixels to ensure the boundary smoothness. In order to minimize (1), we use the efficient graph cut algorithm [Boykov and Kolmogorov 2004] to obtain the final pixel-wise sky/non-sky label assignments. However, the sky segments often contain fine details (e.g., small sky patches among tree regions) for the problem considered in this work. As such, we assign soft labels around the boundary with alpha mattes [Shahrian et al. 2013], and obtain final composition results that are visually appealing after replacing the sky.

Sky Segmentation Results. We quantitatively evaluate the sky segmentation results on the LMSun dataset. We use the same training and test sets as for scene parsing and model learning. The average IOU is improved from 87.6% to 88.7% after refining the sky segmentation. In addition, we compute the boundary precision-recall (BPR) [Galasso et al. 2013] to measure the quality of boundaries. For refined sky segmentation results, we obtain the BPR as 0.839

with online models, which significantly outperform the FCN with the BPR of 0.639. These refinements are important since inaccurate boundaries will result in obvious artifacts during the sky replacement step. We also compare our refined results to those generated by the state-of-the-art segmentation method with dense CRF [Chen et al. 2015], whose 87.9% average IOU is lower than that of our method. Figure 5 further presents the distribution of results in terms of IOU ratio and shows that we have more results over 90% IOU, which is usually considered visually pleasing without much need of manual refinements. Figure 4 shows two examples in which the proposed coarse-to-fine method generates accurate sky segmentations, thereby facilitating better composition around the boundaries.

5 Sky Search

In this section, we introduce the proposed algorithm that searches exemplar images for sky replacement, which is of critical importance for generating visually pleasing results. As discussed in Section 3, a sky region from a reference image with similar content to the input image is more suitable for replacement. Existing methods use global descriptors such as GIST features to search for similar images [Hays and Efros 2007; Liu et al. 2014]. However, the GIST descriptors only describe the global scene layout without important semantic information, and are more effective when the reference images are holistically similar to the inputs, thereby limiting its use for generating composite images with diverse styles. We propose a novel semantic search approach to find a reference image R that is similar to the input image I in both the semantic content and spatial layout, yet with diverse sky appearance. In the following, we first illustrate how we compute the semantic scene layout descriptors , and then introduce a sky selection scheme.

Semantic Layout Descriptors. As described in Section 3, we can obtain a response map from the FCN output layer for each category. We first normalize all the response maps to the range from 0 to 1. Given the response $F_i = \{f_i^1, f_i^2, \ldots, f_i^n\}$ for each pixel i with n categories (i.e., $n = 11$ in this work), its label histogram can be computed as an average pooling process: $H = [h^1; h^2; \ldots; h^n]$, where $h^j = \frac{1}{m}\sum_{i=1}^{m} f_i^j$ and m is the number of pixels. This label histogram indicates the semantic distribution of the region of interest. To obtain a structural version of this descriptor, we use the spatial pyramid pooling method as described in [Lazebnik et al. 2006]. First, we divide the image into three by three grids, and extract histograms H_s for grid s (i.e., $s = 9$ here). Second, a global histogram from the entire image is extracted. After concatenating all the histograms, a final descriptor is constructed as: $\mathbf{H} = [H_1; H_2; \ldots; H_{s+1}]$, where the dimension of $\mathbf{H}$ to describe each image is $n * (s + 1)$.

These descriptors capture both spatial information and semantic cues. Note that [Yan et al. 2016] develops a contextual feature descriptor based on a semantic label map generated by traditional scene parsing [Tighe and Lazebnik 2013] and object detection [Wang et al. 2013] approaches. Their descriptors adopt a fine-grained pooling scheme at multiple scales around each pixel, due to the need of enabling CNN training for local photo adjustment. In contrast, we aim to better capture the overall semantic layout of the image and allow certain degree of layout variations during image search. Therefore, our semantic layout descriptor is designed to be a flexible feature representation with spatial pyramid pooling over the entire response maps, and is constructed on the semantic distribution of all the possible labels, instead of on the final semantic label map as in [Yan et al. 2016]. As demonstrated in Section 7, these descriptors are effective for finding relevant reference sky images for both replacement and appearance transfer. Figure 8 shows the composite results based on retrieved reference sky images by the proposed semantic search method and descriptors, the GIST based approach and random selection. The results indicate that our method can generate more realistic images with diverse styles.

Selection of Sky Images. In addition to semantic matching, the retrieved images need to be further pruned by a few properties, including aspect ratio and resolution, to ensure that the replaced sky can align well in the target image.

Although the sky regions are less sensitive to scale changes, we consider the aspect ratio and resolution to ensure that the reference images are not significantly deformed or distorted for alignment. We compute the aspect ratio $P_a = \frac{\text{width}}{\text{height}}$ and resolution $P_s = \text{width} * \text{height}$ for each sky region. Then a metric [Tsai et al. 2015] comparing I^{sky} and R^{sky} is computed as $Q = \frac{\min(P^I, P^R)}{\max(P^I, P^R)}$, where P^I and P^R are properties for original and reference skies, and Q can be defined for aspect ratio or resolution (i.e., Q_a or Q_s). Note that each measurement is between 0 and 1 and a threshold (i.e., 0.5) is applied to determine whether the sky can be used for replacement or not. If any of the above conditions is not satisfied, we evaluate the next retrieved sky image for replacement.

Diversity of Sky Images. One of our goals is to automatically replace the skies with diverse stylized backgrounds. In order to ensure diversity in the retrieved images, we select sky images based on the inner product of color histograms between reference skies. In addition, this step also enables the flexibility of our system. For instance, if a user prefers strong diversity, the system rejects sky images with a high color similarity in comparison to the ones already selected. On the other hand, if the user prefers a certain sky style, we set a color similarity close to the preferred sky.

6 Sky Replacement

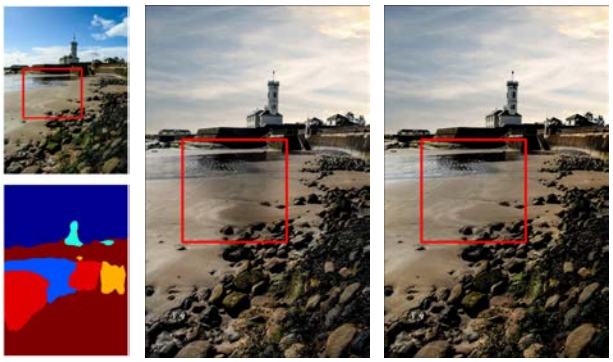
Given a selected sky region R^{sky}, we align and place it in the input image I for replacement. We first extract the maximum rectangular sky region within R^{sky}, and re-scale this extracted sky rectangle to the size of the minimum rectangle that covers all the sky regions of the input image. Since the necessary scale and aspect ratio changes have been addressed in the search step as described in Section 5, the selected new sky region would not have significant distortion and its main interesting region would be retained.

After compositing the new sky R^{sky} into the input image, color adjustment of the foreground region I^{fg} is required to make it compatible with R^{sky}. As a reference image shares similar semantic content with the input image, we propose a semantic-aware transfer method to adjust foreground appearance.

6.1 Semantic-aware Transfer

Transferring color statistics from one image to another is a common technique in image editing. Existing approaches usually perform transfer over the entire region without taking visual semantics into account [Reinhard et al. 2001; Tao et al. 2009] and generate less realistic results when the image content are not well matched. If we take an image pair with beach and sea images for example, without knowing the content, the color from the blue sea may be transferred to a white sand beach, thereby rendering unnatural bluish sand regions. A few local transfer approaches have been developed to directly transfer one region to another [Wu et al. 2013; Laffont et al. 2014]. However, due to large appearance variation between different local regions, the resulting images usually contain artifacts around boundaries, which are difficult to be removed by post-processing (e.g., bilateral filtering).

In this work, we exploit visual semantics based on our scene parsing



(a) Input (b) Direct local transfer (c) Our soft mapping

Figure 6: *(a) Input image (before replacing the sky) and its scene parsing result. (b) After applying the local transfer functions and filtering [Laffont et al. 2014], there are artifacts around boundaries. (c) The proposed method generates smooth result before applying any filter.* *Photo credit: gordon.milligan.*

results for transferring color tones. More importantly, we propose a simple yet effective method that uses soft mapping between semantic regions to generate results with smooth boundaries (see Figure 6 for comparison). Suppose the total number of semantic labels existing in the scene parsing map of the input image is n_r, we formulate the transfer process for each pixel x as:

$$T(x) = \sum_{n=1}^{n_r} W_n(x) \cdot T_n(x), \tag{2}$$

where $W_n(x)$ is the likelihood value in the normalized FCN response map on pixel x for semantic category n, and $\sum_n W_n(x) = 1$. For each semantic label n in the scene parsing results of the input image, we compute a category-specific color/luminance transfer function T_n (defined in Section 6.2) using the regions associated with this label in the input and reference image. Intuitively, local transferring on a pixel x can be interpreted as a soft interpolation according to its semantic responses. The more likely pixel x belongs to label n, the more it would rely on transfer function T_n. This soft mapping based method can largely mitigate the errors in scene parsing (Figure 6), and generate realistic results with smooth boundary transitions, as shown in Figure 11.

Implementation Details. In practice, there may be some small noisy responses from irrelevant labels for each pixel x. As such, we only retain the labels with the top few responses for interpolation, and re-normalize the weights with unity sum. We also remove the foreground object label when generating W_n, as we have merged objects of different categories into this label during the scene parsing, and the appearance variation within this label is too large to have any semantic consistency. After applying transfer functions, we use the original image as guidance, and apply the guided filter [He et al. 2013] to make it better aligned with image edges.

6.2 Transfer Functions

In this section we describe the details of the category-specific transfer function T_n in (2) for each semantic label in the input image. For a semantic label n, even though the reference and input images are semantically similar, it is still possible that there are no regions assigned with label n in the reference image. Thus, we compute T_n based on whether there are regions associated with label n in both input and reference images or not.

Matched region. In the first case, suppose I_n and R_n are the regions associated with the label n in the input and reference images respectively, we then compute both luminance and chrominance transfer functions from R_n to I_n. For luminance, we shift the mean of luminance (L channel of the LAB color space) in I_n to the one in R_n. In addition, we observe that the foreground appearance should not change much when the new sky is similar to the original one, while the appearance may change drastically with a differently stylized sky image. Hence, we compute color differences between the original sky and the new one, and use it to regularize the shift of luminance mean. Specifically, suppose $c(I^{sky})$ and $c(R^{sky})$ are the means of color in I^{sky} and R^{sky} respectively, we have $\beta = \tanh(\left|c(I^{sky}) - c(R^{sky})\right|)$. Then we compute the new desired mean of luminance as: $\hat{L} = L(I_n) + \beta(L(R_n) - L(I_n))$, where $L(R_n)$ and $L(I_n)$ denote the means of luminance in R_n and I_n. It indicates that when the original and reference skies are significantly different, we aggressively adjust more luminance to match R_n, and when the skies are similar, we retain the luminance of I_n.

For chrominance transfer, we edit the channels in the LAB color space by matching the mean and covariance of I_n to R_n using the regularized matching method of [Lee et al. 2016], which performs robustly in practice.

Non-matched region. In the second case, when there is no matched region found in the reference image for label n, we resort to the entire foreground region. That is, we compute a transfer function from the entire reference foreground to the entire original foreground, and use it to represent T_n. We use the same method described in the first case for transferring luminance. For color adjustment, the visual results are more sensitive since no semantic matching is enforced between two foreground regions. Therefore, we transfer the color temperature (CCT) in the XYZ color space [Xue et al. 2012] rather than the chrominance, which is more conservative but robust to semantic inconsistencies.

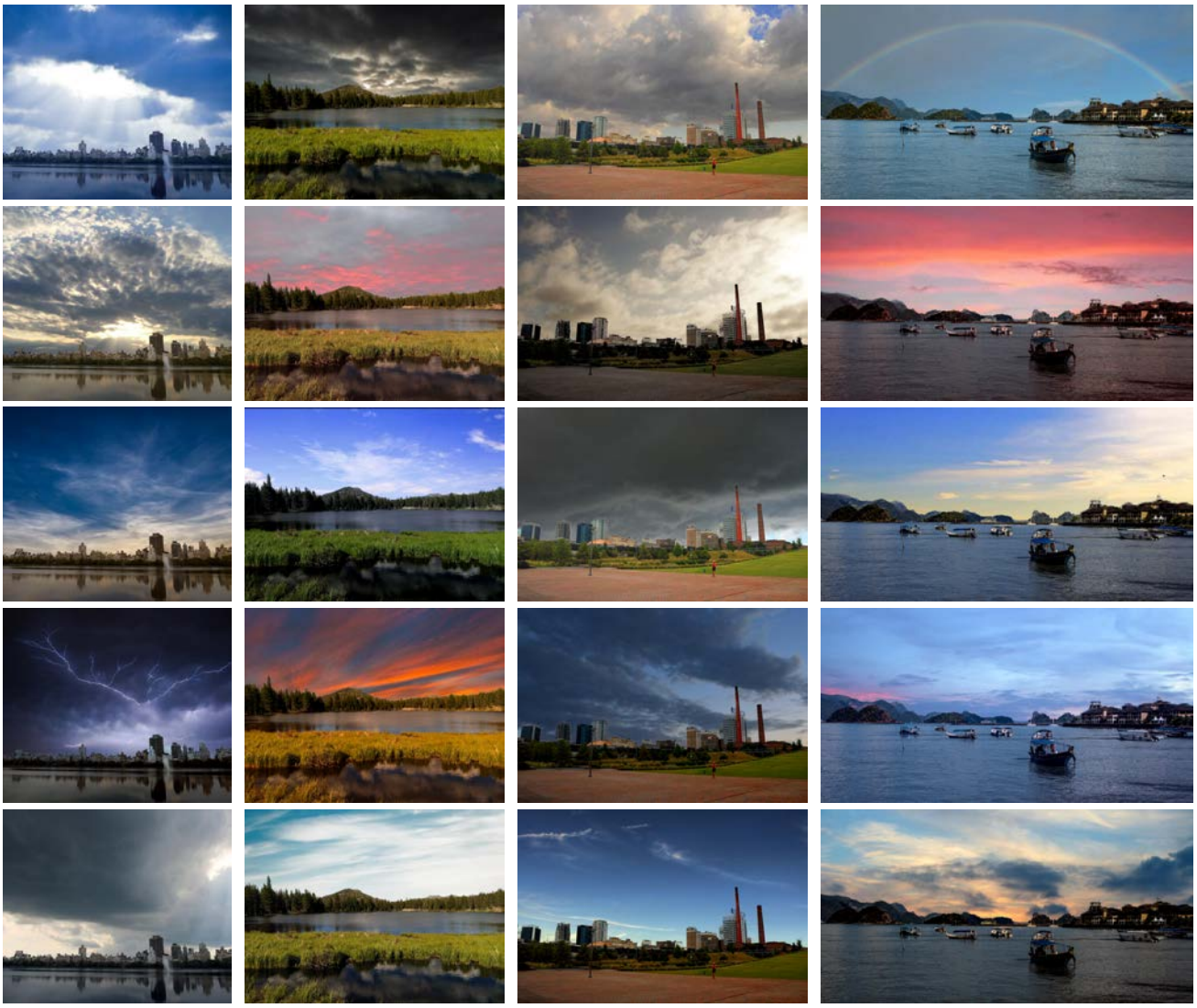
To further prevent from generating artifacts due to inconsistent matching, we use a continuous transfer function for histogram matching in a way similar to the regularization used in [Lee et al. 2016]. Please refer to the supplemental material for details.

7 Results and Analysis

We evaluate the proposed algorithm for rendering composite images with stylized sky regions using a large set from our own collection and Flickr. Figure 7 shows a subset of the generated composite images, and more results and comparisons are provided in the supplementary material. Experimental results demonstrate that our algorithm can handle input images with a variety of scenes and generate a diverse set of visually pleasing sky backgrounds. To quantitatively evaluate the quality of the proposed method, we randomly select 30 test images for user studies. We design three different tasks to evaluate different components of the proposed algorithm. These tasks include comparisons of the proposed sky search and transfer methods to baseline and existing approaches, as well as the comparison of realism of rendered images by the proposed method with respect to the input photographs.

Comparison of sky search methods. We first compare our semantic search method with random selection and the GIST based retrieval approach [Hays and Efros 2007; Liu et al. 2014]. Note that the same semantic-aware transfer method is used for appearance adjustment for all three methods. Qualitatively, the sky examples retrieved by random selection usually do not match the input images well. On the other hand, the method based on the GIST descriptors does not always find images with similar layout and foreground regions as visual semantics are not exploited. In contrast, the proposed method retrieves reference images with diverse styles. Figure 8 shows an example comparing those methods.

(a) Input image



(b) Composite images by the proposed algorithm

Figure 7: *Composite images with stylized sky backgrounds generated by the proposed algorithm. Given an input image (a), we show the top five results (b) with a set of composite images with diverse sky backgrounds.* *Photo credits: Guillermo Palacios, Pat Hawks, Max Wolfe, Tatiana Vdb, amira_a, Jonathan Combe, Miguel Ángel Garca., Tim Green, Seabamirum, Rachel Kramer, PokoPoko, David Stanley, Elescir, Manu Manohar, muffinn, Nisha A, Vtaliy, Robert Couse-Baker, YiKuen Tsai, Hans Kylberg, David Smith, Patrick Metzdorf, David DeHetre and Dilexa.*

To better understand the performance of these methods, we perform user studies for quantitative evaluations. Participants were shown an input image and the top five results generated by the three methods. Each subject is asked to rate each set based on how interesting and realistic the images are, using 5-point Likert scale (1 being worst and 5 best). A set of 1028 scores from 39 subjects is tallied. The proposed method obtains the best average score of 3.42, while the average scores are 3.17 for the GIST based method and 3.18 for random selection. The scores of individual images in the evaluation set are shown in Figure 9, and our method almost always achieves scores larger than 3 while the other two often fail with low scores.

Comparison of sky transfer methods. Next, we compare the proposed semantic-aware sky transfer method to the SkyFinder approach [Tao et al. 2009] which matches the mean and standard deviation in the LAB space, and a baseline transfer method without

(a) Input image (b) Random selection (c) GIST based method (d) Our search method

Figure 8: *An example of the comparison for different sky searching method. For each method, we search the top four skies to replace the input image (a), and use the same technique to transfer appearance. In (b), random selection may find arbitrary skies that are not proper to match statistics. GIST based method (c) is able to find a diverse set of skies, but it may produce non-realistic results with artifacts due to the poor matching between images (e.g., the third and fourth results). Our semantic search approach (d) can handle the both issues, and produces a set of results with diverse skies that are visually pleasing.* *Photo credits: Eoin O'Mahony, Cytryna, Glenn Beltz, Falk Lademann, Jonathan Combe, Otávio Nogueira, Rick Seidel, Ben Laufer, jar [], Robert Couse-Baker, Johan Larsson, YiKuen Tsai and PokoPoko.*

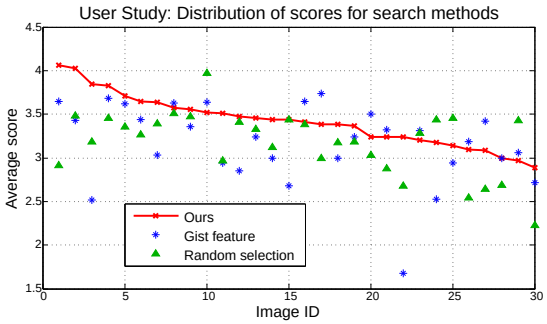


Figure 9: *Average scores of three different methods for each image with x-axis sorted by our score. The proposed technique outperforms random selection in 80% and GIST in 63% of cases. Even in a few cases that our method does not perform well, the results are close to those by the other two schemes.*

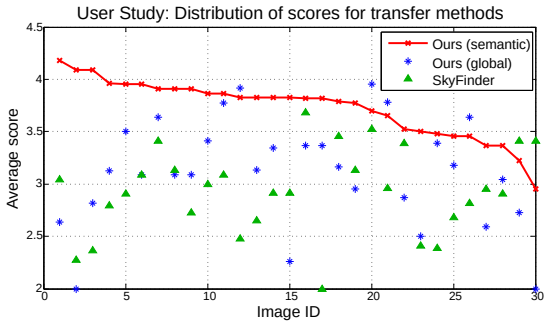
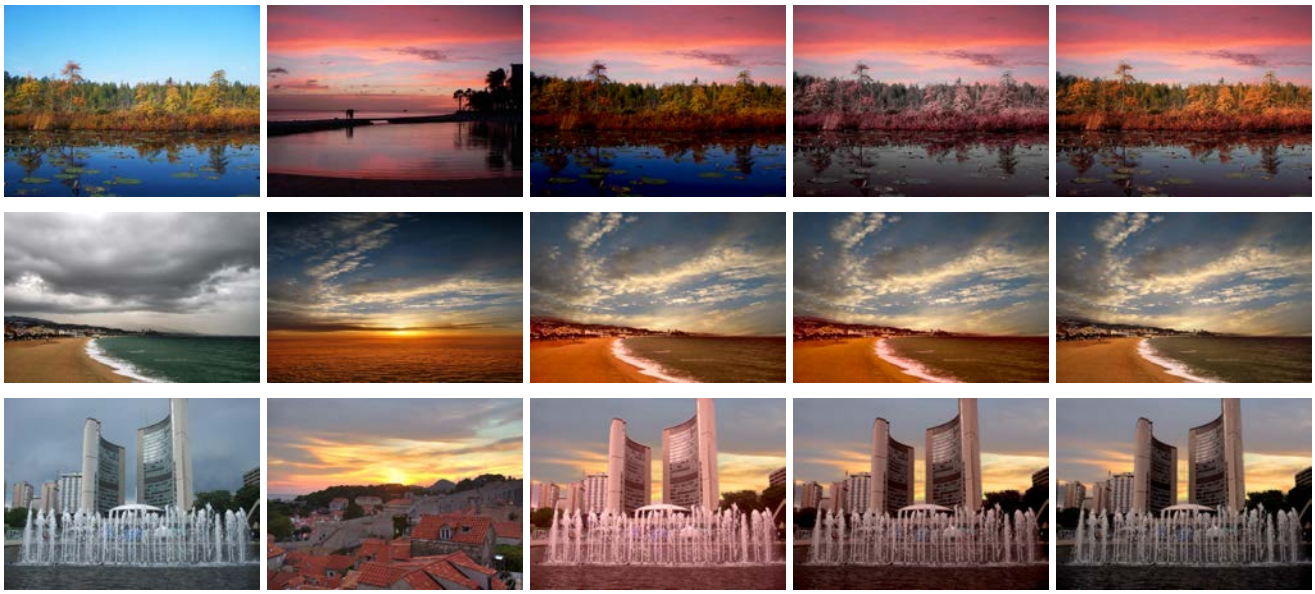


Figure 10: *Average evaluation scores for each image, sorted by our score. Overall, the proposed technique performs better than other two methods in most cases.*

using semantic cues (i.e., directly match from R^{fg} to I^{fg}). The first two methods compute the transfer functions based on the entire foreground regions, which usually generate results with obvious artifacts or over-colorized and unrealistic foregrounds. In contrast, our method takes semantic cues into account and matches regions locally, while using a soft mapping strategy to reduce artifacts around boundaries. Figure 11 shows sample results generated by these three methods.

We conduct user studies to evaluate these methods using similar setups as the previous experiments. Each subject is asked to rate each set based on how realistic the images are. From 27 subjects, we obtain 627 scores for evaluation on transfer methods. On average, the proposed semantic transfer approach achieves the best score of 3.73, while the average score is 3.11 for the baseline transfer method without using semantic information. In contrast, the users give an average score of 2.93 for the SkyFinder method. Average scores of individual test images are shown in Figure 10 where most of our results are rated above 3.5 and are significantly higher than the other two methods.

Comparison to real photographs. Our third user study is designed to evaluate the visual realism and interestingness of the rendered results compared to original (real) photographs. In this study, each participant evaluates a set of image pairs, where each one contains an original image and one of our rendered results that has the same foreground and a stylized sky background. We randomly choose one of our five results for comparisons to ensure that each user only sees an example test image at a time. Specifically, each user is shown with the input and one rendered result side by side in randomized order, and is asked to select the more realistic image.



(a) Input image (b) Reference image (c) [Tao et al. 2009] (d) Ours (w/o semantic) (e) Ours (semantic)

Figure 11: *Rendered results by different sky transfer methods. Given the input image (a) and reference image (b), we show the results using the transfer method proposed in the SkyFinder method [Tao et al. 2009] (c), our method without using semantic matching (d) and our semantic transfer approach (e). For the global methods (c) and (d), the results are likely to contain clear artifacts (top row), over-colorized and unnatural foreground regions (middle and bottom row) due to transfer methods and reference images. In contrast, our method is robust to the reference images and can generate photorealistic results.* *Photo credits: Andy Arthur, mark.watmough, Jorge Franganillo, David Stanley, Vince Alongi and Danny Molyneux.*

A total of 40 subjects participate in this study and a total of 1054 results are tallied. Overall, 61.9% of the real images are favored over the rendered results by the proposed algorithm, 21.2% are rated equally realistic between the two, and interestingly in 16.9% of the test cases our results are considered more realistic than the original images. Since the image pairs are presented side by side, it is easy to find small artifacts in the edited results by directly comparing with the original images, yet still in 38.1% of cases, our results are rated no worse than the real photographs. It indicates the proposed algorithm is able to generate visually pleasing images despite the challenging test set.

In addition, we also ask the subjects to select the image that is more interesting. Among all the evaluated images, 51.2% prefer our results, and 16.1% consider both are equally interesting, while only 32.7% of the cases are in favor of the original images. These results indicate that our algorithm is able to generate more appealing images with interesting styles than the original ones in most cases. For all user studies, the p values are smaller than 0.001, showing the results are statistically significant.

Sky Replacement with Relevance Feedback. In addition to generating images with stylized backgrounds automatically, our system can also be used to guide a user to find preferred sky styles by combining semantic and visual search. Once a user selects a preferred style from our initial results, our system can generate more similar images for users to further explore in a fine-grained manner.

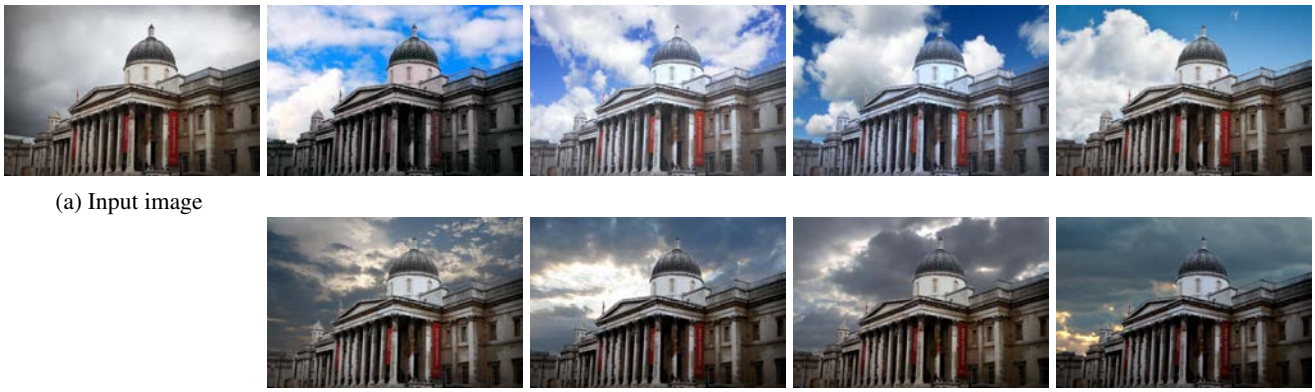
To achieve this, similar to the sky search method described in Section 5, we use our semantic search approach to first rank images in the database to ensure the consistency in image content. We then compute the color similarity between the preferred sky by a user and ranked database images, where this similarity can be set flexibly to control the diversity of retrieved sky backgrounds. Figure 12 illustrates two examples that similar skies are retrieved when a user preferred reference sky image is given. This relevance-feedback scheme facilitates users to find preferred stylized sky backgrounds, while ensuring the quality and realism of rendered images.

Runtime Performance. We measure the runtime of the proposed algorithm on a desktop computer with 3.4GHz Core Xeon CPU, and normalize all the images with the maximum width or height equal to 800 pixels. Implemented in MATLAB, it takes 12 seconds (0.1 seconds with a Titan X GPU and 12GB memory) for scene parsing and generating the FCN semantic responses, and 4 seconds to refine the segmentation results. In addition, it takes 0.5 seconds to retrieve a sky image, and 4 seconds to match a region (where there are usually 2 to 5 regions in an input image) with the C++ implementation. The runtime performance can be improved with high-performance programming languages and code optimization.

Limitation. While the proposed algorithm considers scene semantics in the sky replacement process, it does not take lighting conditions into account. As a result, it is less effective for images with strong directional lighting or high-level cues like shadow directions and reflections. Figure 13 shows one example where the proposed method does not perform well. One solution to address these issues is to estimate the sunlight direction and perform shadow detection [Lalonde et al. 2011]. Such information can be used as prior during the sky search step to ensure that images with similar sunlight directions are retrieved, which will be addressed in our future work.

8 Conclusion

In this work, we propose an automatic method that utilizes semantic information for rendering images with stylized sky backgrounds. We present an accurate sky segmentation algorithm that is effective in delineating boundaries between foreground and background regions. To find proper images for replacement, we construct a new



(a) Input image

(b) Preferred sky style

(c) Results

Figure 12: *Results of appearance-guided sky replacement. Given the input image (a) and one preferred sky style (b), our system is able to find other similar skies (c) in the database. We show two sets of sky replacement results in each row, where each set includes the result of preferred sky style and the other three results that have the similar sky appearance to the preferred one.* *Photo credits: Shimelle Laine, Ben Laufer, The Algerian, Barbara Walsh, Jeff P, Tony Alter, tsuna72, walmarc04 and Mary Bliss.*

(a) Input image (b) Our result

Figure 13: *An example showing the limitation of our method. Without knowing the strong light source, our method is not able to remove the reflection area.* *Photo credits: Michael Caven and skyseeker.*

database that contains various scenes and skies, and search images with similar semantic content. During the sky replacement step, we show that our semantic-aware transfer method can generate realistic results compared to existing approaches. We exploit semantic information through each component of this work, and show that it facilitates the rendering process. For future work, it is of great interest to explore how to utilize semantic cues for image editing problems such as scene completion or photo re-coloring.

Acknowledgments

This work is supported in part by the NSF CAREER Grant #1149783, NSF IIS Grant #1152576, and a gift from Adobe. Portions of this work were performed while Y.-H. Tsai was an intern at Adobe Research. We thank Flickr users mentioned in the manuscript whose photos are under Creative Commons Attribution License: https://creativecommons.org/licenses/by/2.0/.

References

BITOUK, D., KUMAR, N., DHILLON, S., BELHUMEUR, P., AND NAYAR, S. K. 2008. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph. (proc. SIGGRAPH) 27*, 3.

BOYKOV, Y., AND KOLMOGOROV, V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 1124–1137.

CHEN, L.-C., PAPANDREOU, G., KOKKINOS, I., MURPHY, K., AND YUILLE, A. L. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*.

DALE, K., JOHNSON, M. K., SUNKAVALLI, K., MATUSIK, W., AND PFISTER, H. 2009. Image restoration using online photo collections. In *ICCV*.

DARABI, S., SHECHTMAN, E., BARNES, C., GOLDMAN, D. B., AND SEN, P. 2012. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Trans. Graph. (proc. SIGGRAPH) 31*, 4.

GALASSO, F., NAGARAJA, N., CARDENAS, T., BROX, T., AND SCHIELE, B. 2013. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*.

HACOHEN, Y., SHECHTMAN, E., GOLDMAN, D. B., AND LISCHINSKI, D. 2011. Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph. (proc. SIGGRAPH) 30*, 4.

HAYS, J., AND EFROS, A. A. 2007. Scene completion using millions of photographs. *ACM Trans. Graph. (proc. SIGGRAPH) 26*, 3.

HE, K., SUN, J., AND TANG, X. 2013. Guided image filtering. *PAMI 35*, 6, 1397–1409.

HOIEM, D., EFROS, A. A., AND HEBERT, M. 2007. Recovering surface layout from an image. *IJCV 75*, 1.

JOHNSON, M. K., DALE, K., AVIDAN, S., PFISTER, H., FREEMAN, W. T., AND MATUSIK, W. 2011. Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Trans. Vis. Comp. Graph. 17*, 9.

KAUFMAN, L., LISCHINSKI, D., AND WERMAN, M. 2012. Content-aware automatic photo enhancement. *Comp. Graph. Forum 31*, 8.

KUMAR, M. P., TORR, P., AND ZISSERMAN, A. 2005. Obj cut. In *CVPR*.

LAFFONT, P.-Y., REN, Z., TAO, X., QIAN, C., AND HAYS, J. 2014. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Trans. Graph. (proc. SIGGRAPH) 33*, 4.

LALONDE, J.-F., AND EFROS, A. A. 2007. Using color compatibility for assessing image realism. In *ICCV*.

LALONDE, J.-F., HOIEM, D., EFROS, A. A., ROTHER, C., WINN, J., AND CRIMINISI, A. 2007. Photo clip art. *ACM Trans. Graph. (proc. SIGGRAPH) 26*, 3.

LALONDE, J.-F., NARASIMHAN, S. G., AND EFROS, A. A. 2010. What do the sun and the sky tell us about the camera? *IJCV 88*, 1.

LALONDE, J.-F., EFROS, A. A., AND NARASIMHAN, S. G. 2011. Estimating the natural illumination conditions from a single outdoor image. *IJCV 98*, 2, 123–145.

LAZEBNIK, S., SCHMID, C., AND PONCE, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

LEE, J.-Y., SUNKAVALLI, K., LIN, Z., SHEN, X., AND KWEON, I. S. 2016. Automatic content-aware color and tone stylization. In *CVPR*.

LIU, Y., AND YU, Y. 2012. Interactive image segmentation based on level sets of probabilities. *IEEE Transactions on Visualization and Computer Graphics 18*, 2, 202–213.

LIU, C., YUEN, J., AND TORRALBA, A. 2011. Nonparametric scene parsing via label transfer. *PAMI 33*, 12, 2368–2382.

LIU, Y., COHEN, M., UYTTENDAELE, M., AND RUSINKIEWICZ, S. 2014. Autostyle: Automatic style transfer from image collections to users image. *Comp. Graph. Forum 33*, 4.

LONG, J., SHELHAMER, E., AND DARRELL, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

PÉREZ, P., GANGNET, M., AND BLAKE, A. 2003. Poisson image editing. *ACM Trans. Graph. (proc. SIGGRAPH) 22*, 3.

PITIÉ, F., AND KOKARAM, A. 2007. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *CVMP*.

REINHARD, E., ASHIKHMIN, M., GOOCH, B., AND SHIRLEY, P. 2001. Color transfer between images. *IEEE Comp. Graph. Appl. 21*, 5, 34–41.

ROTHER, C., KOLMOGOROV, V., AND BLAKE, A. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (proc. SIGGRAPH) 23*, 3.

SHAHRIAN, E., RAJAN, D., PRICE, B., AND COHEN, S. 2013. Improving image matting using comprehensive sampling sets. In *CVPR*.

SHIH, Y., PARIS, S., DURAND, F., AND FREEMAN, W. T. 2013. Data-driven hallucination of different times of day from a single outdoorphoto. *ACM Trans. Graph. (proc. SIGGRAPH Asia) 32*, 6.

SUNKAVALLI, K., JOHNSON, M. K., MATUSIK, W., AND PFISTER, H. 2010. Multi-scale image harmonization. *ACM Trans. Graph. (proc. SIGGRAPH) 29*, 4.

TAI, Y.-W., JIA, J., AND TANG, C.-K. 2005. Local color transfer via probabilistic segmentation by expectation-maximization. In *CVPR*.

TAO, L., YUAN, L., AND SUN, J. 2009. Skyfinder: Attribute-based sky image search. *ACM Trans. Graph. (proc. SIGGRAPH) 28*, 3.

TAO, M. W., JOHNSON, M. K., AND PARIS, S. 2013. Error-tolerant image compositing. *IJCV 103*, 2, 178–189.

TIGHE, J., AND LAZEBNIK, S. 2013. Superparsing: Scalable nonparametric image parsing with superpixels. *IJCV 101*, 2, 329–349.

TORRALBA, A., OLIVA, A., CASTELHANO, M., AND HENDERSO, J. M. 2006. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review 113*, 10, 766–786.

TSAI, Y.-H., HAMSICI, O., AND YANG, M.-H. 2015. Adaptive region pooling for object detection. In *CVPR*.

WANG, X., YANG, M., ZHU, S., AND LIN, Y. 2013. Regionlets for generic object detection. In *ICCV*.

WU, F., DONG, W., KNOG, Y., MEI, X., PAUL, J.-C., AND ZHANG, X. 2013. Content-based coulour transfer. *Comp. Graph. Forum 32*, 1.

XUE, S., AGARWALA, A., DORSEY, J., AND RUSHMEIER, H. 2012. Understanding and improving the realism of image composites. *ACM Trans. Graph. (proc. SIGGRAPH) 31*, 4.

YAN, Z., ZHANG, H., WANG, B., PARIS, S., AND YU, Y. 2016. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph. 35*, 2.

ZHENG, S., JAYASUMANA, S., ROMERA-PAREDES, B., VINEET, V., SU, Z., DU, D., HUANG, C., AND TORR, P. 2015. Conditional random fields as recurrent neural networks. In *ICCV*.

ZHU, J.-Y., KRÄHENBÜHL, P., SHECHTMAN, E., AND EFROS, A. A. 2015. Learning a discriminative model for the perception of realism in composite images. In *ICCV*.