Manipulating Attributes of Natural Scenes via Hallucination

LEVENT KARACAN, Hacettepe University, Turkey
ZEYNEP AKATA, University of Amsterdam, The Netherlands
AYKUT ERDEM, Hacettepe University, Turkey
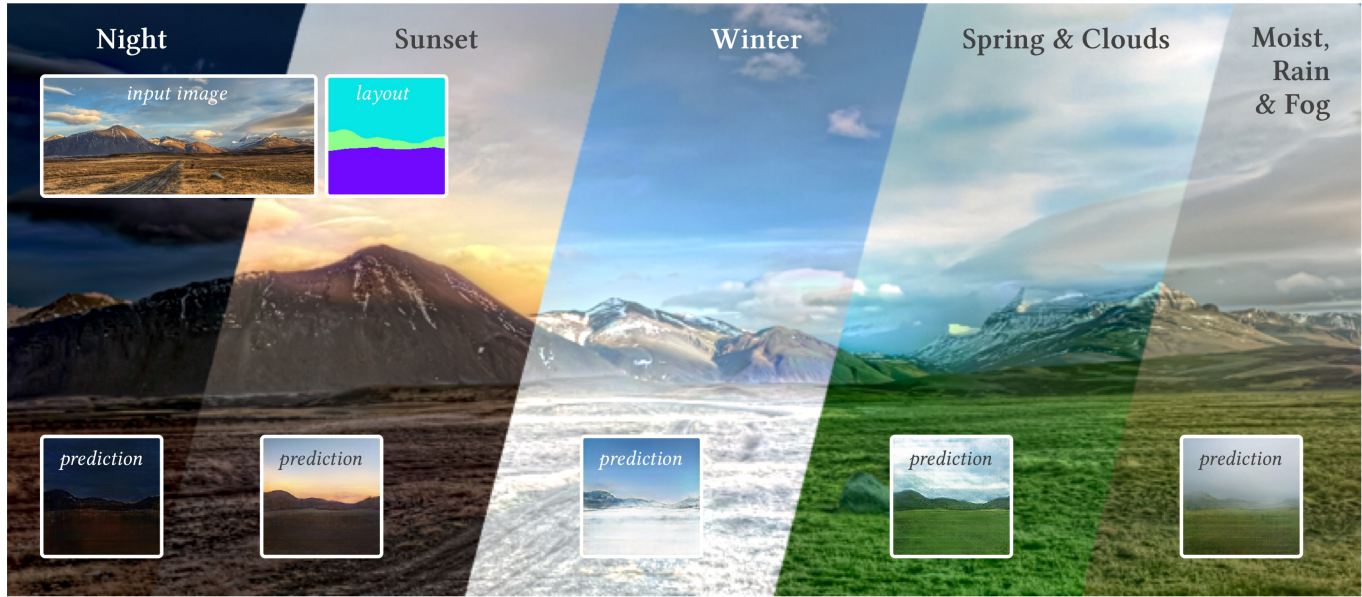ERKUT ERDEM, Hacettepe University, Turkey

Fig. 1. Given a natural image, the proposed approach can hallucinate different versions of the same scene in a wide range of conditions, e.g.*night*, *sunset*, *winter*, *spring*, *rain*, *fog* or even a combination of those. First, our method utilizes a single generator network to imagine the scene with respect to its semantic layout and the desired set of attributes. Then, it directly transfers the look from the hallucinated output to the input image, without a need to have access to a reference style image.

In this study, we explore building a two-stage framework for enabling users to directly manipulate high-level attributes of a natural scene. The key to our approach is a deep generative network which can hallucinate images of a scene as if they were taken at a different season (e.g. during winter), weather condition (e.g. in a cloudy day) or time of the day (e.g. at sunset). Once the scene is hallucinated with the given attributes, the corresponding look is then transferred to the input image while preserving the semantic details intact, giving a photo-realistic manipulation result. As the proposed framework hallucinates what the scene will look like, it does not require any reference style image as commonly utilized in most of the appearance or style transfer approaches. Moreover, it allows to simultaneously manipulate a given scene according to a diverse set of transient attributes within a single model, eliminating the need of training multiple networks per each translation task. Our comprehensive set of qualitative and quantitative results demonstrate the effectiveness of our approach against the competing methods.

CCS Concepts: • **High-Level Image Editing** → **Image Processing**; • **Generative Adversarial Networks** → *Machine Learning*;

Authors' addresses: Levent Karacan, Hacettepe University, Ankara, Turkey, karacan@ cs.hacettepe.edu.tr; Zeynep Akata, University of Amsterdam, Amsterdam, The Netherlands, z.akata@uva.nl; Aykut Erdem, Hacettepe University, Ankara, Turkey, aykut@ cs.hacettepe.edu.tr; Erkut Erdem, Hacettepe University, Ankara, Turkey, erkut@cs. hacettepe.edu.tr.

Additional Key Words and Phrases: Image generation, style transfer, generative models, visual attributes

1 INTRODUCTION

> "The trees, being partly covered with snow, were outlined indistinctly against the grayish background formed by a cloudy sky, barely whitened by the moon."
> – Honore de Balzac (Sarrasine, 1831)

The visual world we live in constantly changes its appearance depending on time and seasons. For example, at sunset, the sun gets close to the horizon gives the sky a pleasant red tint, with the advent of warm summer, the green tones on the grass leave its place in bright yellowish tones and autumn brings a variety of shades of brown and yellow to the trees. Such visual changes in the nature continues in various forms at almost any moment with the effect of time, weather and season. Such high-level changes are referred to as *transient scene attributes* – e.g. cloudy, foggy, night, sunset, winter, summer, to name a few [Laffont et al. 2014].

Recognizing transient attributes of an outdoor image and modifying its content to reflect any changes in these properties were

studied in the past, however, current approaches have many constraints which limit their usability and effectiveness in attribute manipulation. In this paper, we present a framework that can hallucinate different versions of a natural scene given its semantic layout and its desired real valued transient attributes. Our model can generate many possible output images from scratch such as the ones in Figure 1, which is made possible by learning from data the semantic meaning of each transient attribute and the corresponding local and global transformations.

Image generation is quite a challenging task since it needs to have realistic looking outputs. Visual attribute manipulation can be considered a bit harder as it aims at photorealism as well as results that are semantically consistent with the input image. For example, for predicting the look of a scene at sunset, visual appearances of the sky and the ground undergo changes differently, the sky gets different shades of red while the dominant color of the ground becomes much darker and texture details get lost. Unlike recent image synthesis methods [Chen and Koltun 2017; Isola et al. 2017; Wang et al. 2017] which explore producing realistic-looking images from semantic layouts, automatically manipulating visual attributes requires modifying the appearance of an input image while preserving object-specific semantic details in tact. Recent style transfer methods achieve this goal to a certain extent but they require a reference style image. A simple solution to come up with an automatic style transfer method is to retrieve reference style images with desired attributes from a well-prepared dataset with a rich set of attributes. However, this approach raises new issues that need to be solved such as retrieving images according to desired attributes and semantic layout in an effective way. To overcome these obstacles, we propose to combine neural image synthesis and style transfer approaches to perform visual attribute manipulation. For this purpose, we first devise a conditional image synthesis model that is capable of hallucinating desired attributes on synthetically generated scenes with semantic content similar to the input image and then we resort to a photo style transfer method to transfer the visual look of the hallucinated image to the original input image to obtain a resulting image with the desired attributes.

A rich variety of generative models including Generative Adversarial Networks (GANs) [Goodfellow et al. 2014; Radford et al. 2016; Vondrick et al. 2016], Variational Autoencoders (VAEs) [Gregor et al. 2015; Kingma and Welling 2014], and autoregressive models [Mansimov et al. 2016; Oord et al. 2016] are developed to synthesize visually plausible images. Images of higher resolutions, e.g. 128×128 or 256×256, have also been rendered under improved versions of these frameworks [Berthelot et al. 2017; Chen and Koltun 2017; Gulrajani et al. 2016; Reed et al. 2016a,b; Shang et al. 2017; Zhu et al. 2017a]. However, generating diverse, photo-realistic and well-controlled images of complex scenes has not yet been fully solved. For image synthesis, we propose a new conditional GAN based approach to generate target image which have the same semantic layout with the input image but reflects desired transient attributes. Our approach enables users to play with degrees of transient attributes thanks to learned transient attribute manifold, hence we can manipulate multiple attributes of input image just increasing or decreasing related continuous transient attributes as shown in Figure 1.

To build the aforementioned model, we argue the necessity of better control over the generator network in GAN. We address this issue by conditioning ample concrete information of scene contents to the default GAN framework, deriving our proposed attribute and semantic layout conditioned GAN model. Spatial layout information tells the network where to draw, resulting in clearly-defined object boundaries and transient scene attributes serve to edit visual properties of a given scene so that we can hallucinate desired attributes for input image in semantically similar generated image.

However, naively importing the side information is insufficient. For one, when training the discriminator to distinguish mismatched image-condition pairs, if the condition is randomly sampled, it can easily be too off in describing the image to provide meaningful error derivatives. To address this issue, we propose to selectively sample mismatched layouts for a given real image, inspired by the practice of hard negative mining [Wang and Gupta 2015]. For another, given the challenging nature of the scene generation problem, adversarial objective alone can struggle to discover a satisfying output distribution. Existing works in synthesizing complex images apply the technique of "feature matching", or perceptual loss [Chen and Koltun 2017; Dosovitskiy and Brox 2016]. Here, we also adopt perceptual loss to stabilize and improve adversarial training for more photographic generation but contrasting prior works, our approach employs the layout-invariant features pretrained on segmentation task to ensure consistent layouts between synthesized images and reference images. For photo style transfer, we use a recent deep learning based approach [Luan et al. 2017] which transfers visual appearance between same semantic objects in real photos using semantic layout maps.

1.1 Contributions

Our contributions are summarized as follows:

- We propose a new two-stage visual attribute manipulation framework for changing high-level attributes of a given outdoor image.
- We develop a conditional GAN variant for generating natural scenes faithful to given semantic layouts and transient attributes.
- We build up an outdoor scene dataset annotated with layout and transient attribute labels by combining and annotating images from Transient Attributes [Laffont et al. 2014] and ADE20K [Zhou et al. 2017].

Our code and models are publicly available on the project website[1].

2 RELATED WORK

2.1 Image Synthesis

Recently, much progress has been made towards realistic image synthesis; in particular, different flavors and improved versions of Generative Adversarial Networks (GANs) have achieved impressive results along this direction. Radford et al. [2016] establish a state-of-the-art architecture, DCGAN, to enable training on larger scale and higher resolution datasets and many follow-up works attempt to improve upon DCGAN [Arjovsky et al. 2017; Mao et al. 2016;

---

[1]https://web.cs.hacettepe.edu.tr/~karacan/projects/attribute_hallucination/

Salimans et al. 2016]. Larsen et al. [2015] integrates adversarial discriminator to VAE framework in an attempt to prevent mode collapsing and its extension [Shang et al. 2017] further tackles this issue while improving generation quality and resolution.

Conditional GANs (CGANs) [Mirza and Osindero 2014] that leverages side information have been widely adopted to generate images under predefined constraints. For example, Reed et al. [2016a,b] generate images using natural language descriptions; Antipov et al. [2017] follow similar pipelines to edit a given facial appearance based on age. Pix2pix [Isola et al. 2017] undertakes a different approach to conditional generation that it directly translates one type of image information to another type through an encoder-decoder architecture coupled with adversarial loss; its extension Cycle-GAN conducts similar translation under the assumption that well-aligned image pairs are not available. The design of our image synthesis model resembles CGANs, as opposed to Pix2pix, since those so-called image-to-image translation models are limited in terms of output diversity.

In the domain of scene generation, the aforementioned Pix2Pix [Isola et al. 2017] and Cycle-GAN [Zhu et al. 2017a] both manage to translate realistic scene images from semantic layouts. However, these models are deterministic, in other words, they can only map one input image to one output image in different domains. Recently, some works [Huang et al. 2018; Zhu et al. 2017b] have proposed multimodal image to image translation models. Zhu et al. [2017b] has investigated new GAN and VAE based objective functions, network architectures, and methods of injecting the latent code to achieve multimodality. More specifically, they try to provide bijection translation between learned latent code and output image through cycle consistency loss with different noise injection alternatives. Huang et al. [2018] assumes that each image is generated from shared latent code by different domains and domain specific latent code, and trains two encoders between cross domains with adversarial loss to provide domain transition with target domain specific latent code. Although these works improve the diversity to a certain degree, they are still limited in the sense that they do not allow to fully control the latent scene characteristics. For instance, with these methods, it is not possible to obtain an image with half sunset, half cloudy.

Alternatively, some efforts [Chen and Koltun 2017; Wang et al. 2017] on image to image translation has been made to increase the realism and resolution with multi-scale approaches. Chen and Koltun [2017] try to achieve realistic looking scenes through a carefully crafted regression objective that maps a single input layout to multiple potential scene outputs. Nonetheless, despite modeling one-to-many relationships, the number of outputs is pre-defined and fixed, which still puts tight constraints on the generation process. As compared to these works, besides taking semantic layout as input, our proposed scene generation network is additionally aware of the transient attributes and the latent random noises characterizing intrinsic properties of the generated outputs. As a consequence, our model is more flexible in generating the same scene content under different conditions such as lighting, weather, and seasons.

From training point of view, a careful selection of "negative" pairs, i.e. negative mining, is an essential component in metric learning and ranking [Fu et al. 2013; Li et al. 2013; Shrivastava et al. 2016]. Existing works in CGAN have been using randomly sampled negative image-condition pairs [Reed et al. 2016a]. However, such random negative mining strategy has been shown to be inferior to more meticulous negative sampling schemes [Bucher et al. 2016]. Particularly, the negative pair sampling scheme proposed in our work is inspired by the concept of relevant negative [Li et al. 2013], where the negative examples that are visually similar to positive ones are emphasized more during learning.

To make the generated images look more similar to the reference images, a common technique is to consider feature matching which is commonly employed through a perceptual loss [Chen and Koltun 2017; Dosovitskiy and Brox 2016; Johnson et al. 2016]. The perceptual loss in our proposed model distinguishes itself from existing works by matching segmentation invariant features from pre-trained segmentation networks [Zhou et al. 2017], leading to diverse generations that obey the conditioned layouts.

2.2 Image Editing

2.2.1 Visual Appearance Transfer. There has been a great effort towards building methods for manipulating visual appearance of a given image. Example-based approaches [Pitie et al. 2005; Reinhard et al. 2001] use a reference image to transfer color space statistics to input image so that visual appearance of input image looks like the reference image. In contrast of these global color transfer approaches which require highly consistent reference images with input image, user controllable color transfer techniques were also proposed [An and Pellacini 2010; Dale et al. 2009] to consider spatial layouts of input and reference images. Dale et al. [2009] search for some reference images which have similar visual context to input image in a large image dataset to transfer local color from them and then use color transferred image to restore input image. Other local color transfer approaches [Wu et al. 2013] use the semantic segments to transfer color between regions in reference and input images have same semantic label (e.g. color is transferred from sky region in reference image to sky region in input image).

Some data-driven approaches [Laffont et al. 2014; Shih et al. 2013] leverage the time-lapse video datasets taken for same scene to capture scene variations that occur at different times. Shih et al. [2013] aim to give times of day appearances to a given input image, for example converting an input image taken midday to a nice sunset image. They first retrieve the most similar video frame to input scene from dataset as reference frame, then find matching patches between reference frame and input image and lastly transfer the variation occurs between reference frame and desired reference frame which is same scene but taken different time of day to input image. Laffont et al. [2014] take a step forward in their work for handling more general variations as transient attributes such as lighting, weather, and seasons.

2.2.2 High-Level Image Editing. High-level image editing offers easier and more natural way to casual users to manipulate a given image. Instead of using a reference image either provided by the user or retrieved from a database, learning the image manipulations and high-level attributes for image editing like a human has also attracted researchers. Berthouzoz et al. [2011] learn parameters of the basic operations for some manipulations recorded in photoshop as macro to adapt them to new images, for example, applying same

skin color correction operation with same parameters for both faces with dark-skined and light-skined does not give expected correction. In contrast to learning image operations for specific editing effects, Cheng et al. [2014] learn the attributes as adjectives and objects as nouns for semantic parsing of an image and further use them for verbal guided image manipulation to indoor images, for example the verbal command: "*change the floor to wooden*" modifies the appearance of the floor. Similarly, Laffont et al. [2014] learn to recognize transient attributes for image editing on well-defined transient attribute dataset for outdoor scenes using crowd-sourcing and then learned attributes are used to transfer appearance to input image from desired variation of retrieved image which have similar scene features with input image. For example, for an input image taken in a sunny day, suppose a user want this image to look like as if it is taken in a winter day. The method first finds the most similar scene and then winter version of this scene using to predicted attribute labels from the dataset, then it transfers variation between the retrieved scene and its desired version to the input image. Lee et al. [2016] aim to automatically select a subset of style exemplars that will achieve good stylization results by learning a content-to-style mapping between large photo collection and a small style dataset.

2.2.3 Image Editing with Deep Learning. Deep learning has fueled a growing literature on employing neural approaches to improve existing image editing problems. Here, we review the studies that are the most relevant to our work.

Gatys et al. [2016] has demonstrated how Convolutional Neural Networks (CNNs) effectively encode content and texture separately in feature maps of CNNs trained on large-scale image datasets and has proposed a neural style transfer method to transfer artistic styles from paintings to natural images. Alternatively, Johnson et al. [2016] train a transformation network to speed up the test time of style transferring together with minimization of perceptual loss between input image and stylized image. Recent deep photo style transfer method of Luan et al. [2017] aims at providing realism in case of style transfer is made between the real photos. For example, when one wants to make an input photo look like taken in different illumination and weather conditions it is needed a photo-realistic transferring. It uses semantic labels to prevent semantic inconsistency so that style transfer is carried out between same semantic regions. Style transfer networks are also specialized for the editing face images and portraits [Kemelmacher-Shlizerman 2016; Liao et al. 2017; Selim et al. 2016] with new objectives. Style transfer works limit the users to find an reference photo in which desired style effects exist for desired attributes.

Yan et al. [2016] introduce the first automatic photo adjustment framework based on deep neural networks. They use deep neural network to learn a regressor which transforms the colors for artistic styles especially color adjustment from the image and its stylized version pairs. They define a set of feature description based on pixel, global and semantic levels. In another work, Gharbi et al. [2017] propose a new neural network architecture to learn image enhancement transformations at low resolution, then they move learned transformations to higher resolution in bilateral space in an edge-preserving manner.

Lastly, building upon conditional GAN model, some image completion works have been proposed to predict missing regions providing global and local context information with multiple discriminator networks [Iizuka et al. 2017; Li et al. 2017].

3 ATTRIBUTE MANIPULATION FRAMEWORK

Our framework provides an easy and high-level editing system to manipulate transient attributes of outdoor scenes (see Figure 2). The key component of our framework is a scene generation network that is conditioned on semantic layout and continuous-valued vector of transient attributes. This network allows us to generate synthetic scenes consistent with the semantic layout of the input image and having the desired transient attributes. One can play with 40 different transient attributes by increasing or decreasing values of certain dimensions. Note that, at this stage, the semantic layout of the input image should also be fed to the network, which can be easily automated by a scene parsing model. Once an artificial scene with desired properties is generated, we then transfer the look of the hallucinated image to the original input image to achieve attribute manipulation in a photo-realistic manner.

Since our approach depends on a learning-based strategy, it requires a richly annotated training dataset. In Section 3.1, we describe our own dataset, named ALS17K, which we have created for this purpose. In Section 3.2, we present the architectural details of our attribute and layout conditioned scene generation network and the methodologies employed for effectively training our network. Finally, in Section 3.3, we discuss the photo style transfer method that we utilize to transfer the appearance of generated images to the input image. We make our code and dataset publicly available on the project website.

3.1 The ALS17K Dataset

For our dataset, we pick and annotate images from two popular scene datasets, namely ADE20K [Zhou et al. 2017] and Transient Attributes [Laffont et al. 2014], for the reasons which will become clear shortly.

ADE20K [Zhou et al. 2017] includes 22, 210 images from a diverse set of indoor and outdoor scenes which are densely annotated with object and stuff instances from 150 classes. However, it does not include any information about transient attributes. Transient Attributes [Laffont et al. 2014] contains 8, 571 outdoor scene images captured by 101 webcams in which the images of the same scene can exhibit high variance in appearance due to variations in atmospheric conditions caused by weather, time of day, season. The images in this dataset are annotated with 40 transient scene attributes, e.g. sunrise/sunset, cloudy, foggy, autumn, winter, but this time it lacks semantic layout labels.

To establish a richly annotated, large-scale dataset of outdoor images with both transient attribute and layout labels, we further operate on these two datasets as follows. First, from ADE20K, we manually pick the 9,201 images corresponding to outdoor scenes, which contain nature and urban sceneries. For these images, we need to obtain transient attribute annotations. To do so, we conduct initial attribute predictions using the pretrained model from [Baltenberger et al. 2016] and then manually verify the predictions. From Transient Attributes, we select all the 8,571 images. To get the layouts, we
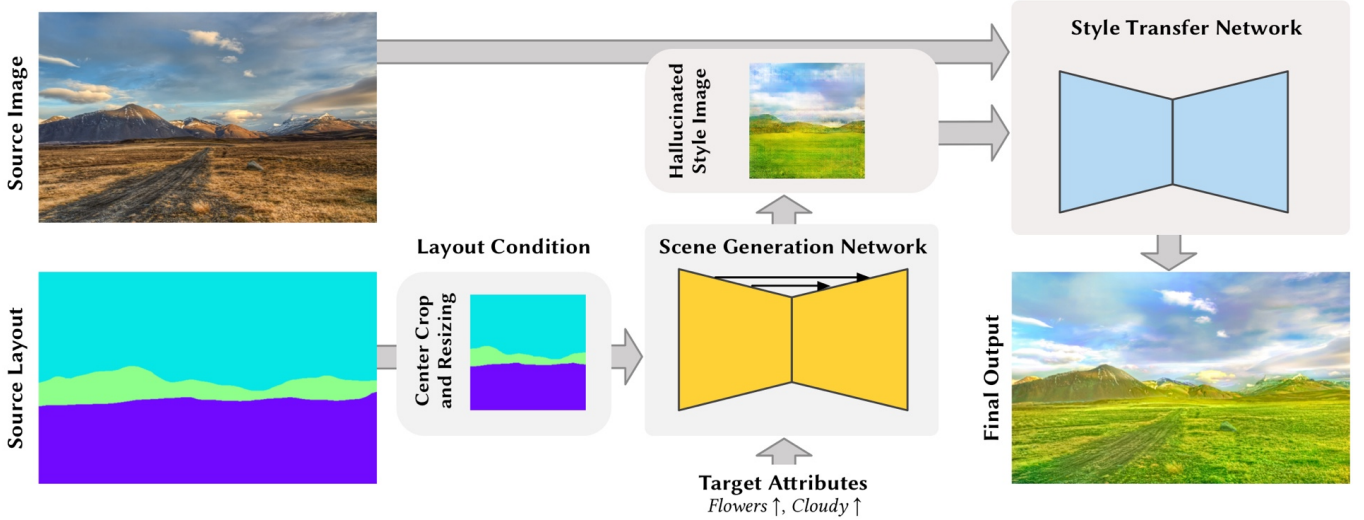
Fig. 2. Overview of the proposed attribute manipulation framework. Given an input image and its semantic layout, we first resize and center crop the layout to 256×256 pixels and feed it to our scene generation network. After obtaining the scene synthesized according to the target transient attributes, we transfer the look of the hallucinated style back to the original input image.

first run the semantic segmentation model in [Zhao et al. 2017], the winner of the MIT Scene Parsing Challenge 2016, and assuming that each webcam image of the same scene has the same semantic layout, we manually select the best semantic layout prediction for each scene and use those predictions as the ground truth layout for the related images.

In total, we collect 17,772 outdoor images (9,201 from ADE20K + 8,571 from Transient Attributes), with 150 semantic categories and 40 transient attributes. Following the train-val split from ADE20K, 8,363 out of the 9,201 images are assigned to the training set, the other 838 testing; for the Transient Attributes dataset, 500 randomly selected images are held out for testing. In total, we have 16,434 training examples and 1,338 testing images. More samples of our annotations are presented in the supplementary materials. Lastly, we resize the height of all images to 256 and then center-crop to 256×256 pixels.

3.2 Scene Generation

In this section, we first give a brief technical summary of GANs and conditional GANs (CGANs), which provides the foundation for our scene generation network (SGN). We then present architectural details of our SGN model, followed by the two strategies applied for improving the training process. All the implementation details are included in the Supplementary Materials.

3.2.1 Generative Adversarial Networks.
Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] have been designed as a two-player min-max game where a discriminator network D learns to determine if an image is real or fake and a generator network G strives to output as realistic images as possible to fool the discriminator. Within this min-max game, G and D can be trained jointly

by performing alternative updates to solve the following objective:

$$\min_{G} \max_{D} V(D, G) \quad = \quad E_{x \sim p_{data}(x)}[\log D(x)] + \qquad (1)$$
$$E_{x \sim p_z(z)}[\log(1 - D(G(z)))]$$

where x is a natural image drawn from the true data distribution $p_{data}(x)$ and z is a random noise vector sampled from a multivariate Gaussian distribution. In [Goodfellow et al. 2014], it is shown that the optimal solution to this min-max game is when the distribution p_G converges to p_{data}.

Conditional GANs [Mirza and Osindero 2014] (CGANs) engage additional forms of side information as generation constraints, e.g. class labels [Mirza and Osindero 2014], image captions [Reed et al. 2016b], bounding boxes and object keypoints [Reed et al. 2016a], and etc. Given a context vector c as side information, the generator $G(z, c)$, taking both the random noise and the side information, tries to synthesize a realistic image that satisfies the condition c. The discriminator, now having real/fake images and context vectors as inputs, aims at not only distinguishing real and fake images but also whether an image satisfies the paired condition c. Such characteristics is referred to as match-aware [Reed et al. 2016b]. In this way, we expect the generated output of CGAN x_g is controlled by the side information c. Particularly, in our model, c is composed of semantic layouts s and transient attributes a.

3.2.2 Proposed Architecture.
Our framework consists of a generator G and a discriminator D, both conditioned on the semantic layout and transient scene attributes. In our model, while the semantic layout categories are encoded into 8-bit binary codes, transient attributes are encoded by a 40-d vector. As illustrated in Figure 3, we concatenate semantic layout S and the spatially replicated attribute vectors a and z, feed their concatenation into G to obtain the final generation. The discriminator D takes in tuples of real or generated images, matching or mismatching semantic layouts and transient
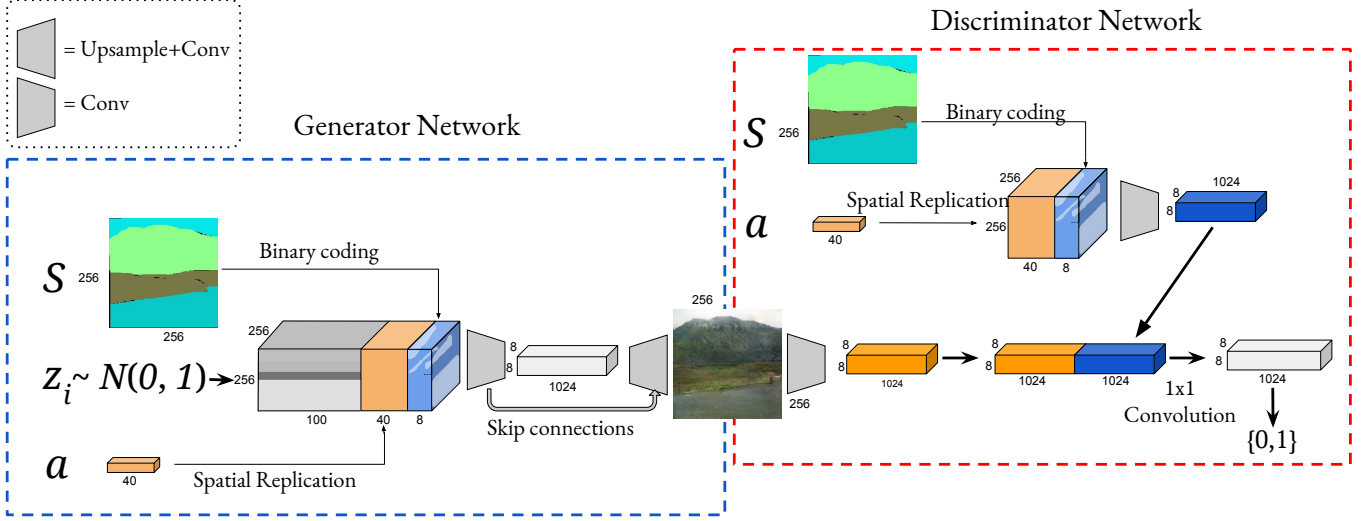
Fig. 3. Scene Generation Network (SGN). Our proposed CGAN architecture for generating synthetic outdoor scenes consistent with given layout and transient attributes.

attributes to decide whether the images are fake or real and whether the pairings are valid. That is

$$D(x, a, s) = \begin{cases} 1, x \in P_{\text{data}} \text{ and } x, a, S \text{ correctly match,} \\ 0, \text{ otherwise.} \end{cases}$$

The overall objective for SGN is thus

$$\min_G \max_D V(D, G) = E_D + E_G \text{ where} \qquad (2)$$

$$E_D = E_{x,S,a \sim p_{data}(x,S,a)}[\log D(x, S, a)] \text{ and}$$

$$E_G = E_{z \sim p_z(z); S, a \sim p_{data}(S,a)}[\log (1 - D(G(z, S, a)))].$$

The generator of SGN shares similar architecture as in DCGAN [Radford et al. 2016]. But instead of fully convolutional layers, we apply nearest neighbor upsampling followed by a convolutional layer [Odena et al. 2016]. The discriminator resembles a Siamese network [Bromley et al. 1994; Chopra et al. 2005], where one stream takes the real/generated image as input, the second one processes the given attribute and the spatial layout labels. The responses of these networks are then concatenated and fused via a 1×1 convolution operation. The combined features are finally sent to fully-connected layers for the binary decision.

3.2.3 Improved Training of SGNs. Here we elaborate on two complementary training techniques that substantially boost the efficiency of the training process.

Relevant Negative Mining. Training the match-aware discriminator in CGAN resembles learning to rank [Rudin and Schapire 2009], in the sense that a "real pair"–real image paired with right conditions–should score higher (i.e. classifying into category 1 in this case) than a "fake pair"–either image is fake or context information is mismatched (i.e. classifying into category 0). For ranking loss, it has been long acknowledged that naively sampling random negative examples is inferior to more carefully designed negative

sampling scheme, such as various versions of hard negative mining [Bucher et al. 2016; Fu et al. 2013; Li et al. 2013; Shrivastava et al. 2016]. Analogously, a better negative mining scheme can be employed by training CGAN, as existing works have been using random sampling [Reed et al. 2016a]. To this end, we propose to apply the concept of relevant negative mining [Li et al. 2013] (RNM) to sample mismatching layout in training our SGN model. Concretely, for each layout S, we search for its nearest neighbor S' and set it as the corresponding mismatching negative example for S. In Section 4, we present empirical qualitative and quantitative results to demonstrate improvement from RNM over random sampling. We attempted similar augmentation on attributes a by flipping a few of them instead of complete random sampling to obtain the mismatching a' but found such operation hurt the performance, likely due to the flipped attributes being too semantically close to the original ones which cause ambiguity to the discriminator.

Layout-Invariant Perceptual Loss. Following the practice of existing works [Chen and Koltun 2017; Dosovitskiy and Brox 2016], we also seek to stabilize adversarial training and enhance generation quality by adding a perceptual loss. Conventionally, features used for perceptual loss come from a deep CNN, such as VGG [Simonyan and Zisserman 2014], pretrained on ImageNet for classification task. However, perceptual loss to match such features would intuitively withhold generation diversity, which opposes our intention of creating stochastic output via a GAN framework. Instead, we propose to employ intermediate features trained on outdoor scene parsing with ADE20K. The reason for doing so is three folds: diversity in generation is not suppressed, because scenes with different contents but the same layout ideally produce the same high-level features; the layout of the generation is further enforced thanks to the nature of the scene parsing network; since the scene parsing network is trained on real images, the perceptual loss will impose additional

---
Algorithm 1: SGN training algorithm

1: **Input:** Training set $\Omega = \{(x, a, S)\}$ with training images x, semantic segmentation layouts S and transient attributes a.
2: **for all** number of iterations **do**
3:    sample minibatch of paired x, a, S
4:    sample minibatch of z_i from $\mathcal{N}(0, I)^Z$
5:    **for all** (x_i, a_i, S_i) in Ω **do**
6:      Randomly sample negative a_i' mismatching x_i
7:      Sample S_i' mismatching x_i via RNM
8:    **end for**
9:    $x_g \leftarrow G(z_i, a_i, S_i)$ {Forward through generator}
10:    $\mathcal{L}_D \leftarrow$
     $-(\log D(x, a, S) + \log(1 - D(x_g, a, S)) + \log(1 - D(x, a', S')))$
11:    $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ {Update discriminator}
12:    $\mathcal{L}_G \leftarrow -\log D(x_g, a, S) + \|f_p(x) - f_p(x_g)\|_2^2$
13:    $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ {Update generator}
14: **end for**
---

regularization to make the output more photo-realistic. The final version of our proposed perceptual loss is as follows:

$$E_P = E_{z \sim p_z(z); x, S, a \sim p_{data}(S, a)}[\|f_P(x) - f_P(G(z, S, a))\|_2^2],$$

where f is the CNN encoder for the scene parser network. We further demonstrate the effectiveness of the proposed layout-invariant perceptual loss in Section 4.

By additionally considering RNM and perceptual loss, we arrive at the training procedure which is outlined in Algorithm 1.

3.3 Style Transfer

The main goal in photo style transfer is to successfully transfer visual style (such as color and texture) of a reference image onto another image while preserving semantic structure of the target image. In the past, statistical color transfer methods [Pitie et al. 2005; Reinhard et al. 2001] showed that the success of the style transfer methods highly depend on the semantic similarity of the source and target images. To overcome this obstacle, user interaction, semantic segmentation approaches or image matching methods were utilized to provide semantic relation between source and target images. In addition, researchers explored data driven methods to come up with fully automatic approaches which retrieve the source style image through some additional information such as attributes, features and semantic similarity.

For existing deep learning based photo style transfer methods, it is still crucial that source and reference images have similar semantic layouts to provide successful and realistic style transfer results. Image retrieval based approaches are limited with the dataset and become infeasible when there is no images with the desired properties. The key distinguishing characteristics of our framework is that we can generate a style image on the fly that has both similar semantic layout with the input image and possess the desired transient attributes, thanks to our proposed SGN model.

In our framework, we employ the recently proposed deep photo style transfer method (DPST) [Luan et al. 2017], which extends the formalization of the neural style transfer method of Gatys et al.

[2016] by adding a photorealism regularization term that enables the style transfer to be done between same semantic regions instead of the whole image. This property makes DPST very appropriate for our image manipulation system. Although this method produces good results, some smoothing effects may hurt the photorealism. For that reason, we additionally apply the post-processing method in [Mechrez et al. 2017] in which screened Poisson equation was used to make the stylized image more similar to the input image in order to increase its visual quality.

4 RESULTS AND COMPARISON

We first evaluate our scene generation network's ability to synthesize diverse and realistic-looking outdoor scenes (Section 4.1). We then show attribute manipulation results of the proposed two-stage framework which employs the hallucinated scenes as reference style images (Section 4.2) and discuss the limitations of the approach (Section 4.3).

4.1 Attribute and Layout Guided Scene Generation

We present a comprehensive set of experiments to assess the effectiveness of our SGN model on generating outdoor scenes in terms of image quality, condition correctness and diversity. We also demonstrate how the proposed model enables the users to add and subtract scene elements. All results reported below are of size 256×256 pixels.

4.1.1 Training Details. We use a setting similar to the one in [Radford et al. 2016]. All models were trained with mini-batch stochastic gradient descent (SGD) with a mini-batch size of 40. Parameters were initialized from a zero-centered Normal distribution with standard deviation of 0.02. We use the Adam optimizer [Kingma and Ba 2014] with the learning rate value of 2×10^{-4} and the momentum value of 0.5. We train our models for 400 epochs on a NVIDIA TITAN X GPU, which lasts about 3 days. Our implementation is based on the Pytorch framework.

4.1.2 Ablation Study. We illustrate the role of Relevant Negative Mining (RNM) and layout-invariant Perceptual Loss (PL) in improving generation quality with an ablation study. The input layouts used in our analysis are taken from the test set, which are unseen during training. Furthermore, we fix the transient attributes to the predictions of the pre-trained deep transient model [Baltenberger et al. 2016]. Figure 4 presents synthetic outdoor images generated from layouts depicting different scene categories such as mountain, forest, coast, lake and highway. We make the following observations from these results.

Attributes of the generated images are mostly in agreement with the original transient attributes. Integrating RNM slightly improves the rendering of attributes but in fact, its main role is to make training more stable. Perceptual loss contributes a lot to boost the final image quality of SGN. The roads, trees and clouds are drawn with the right texture; the color distributions of sky, water, field and etc. also appear realistic; reasonable physical effects are also observed such as the reflection of the water, fading of the horizon, valid view perspective of urban objects. Overall, the results obtained with both RNM and Perceptual Loss are visually more pleasing and faithful to the desired attributes and layouts.
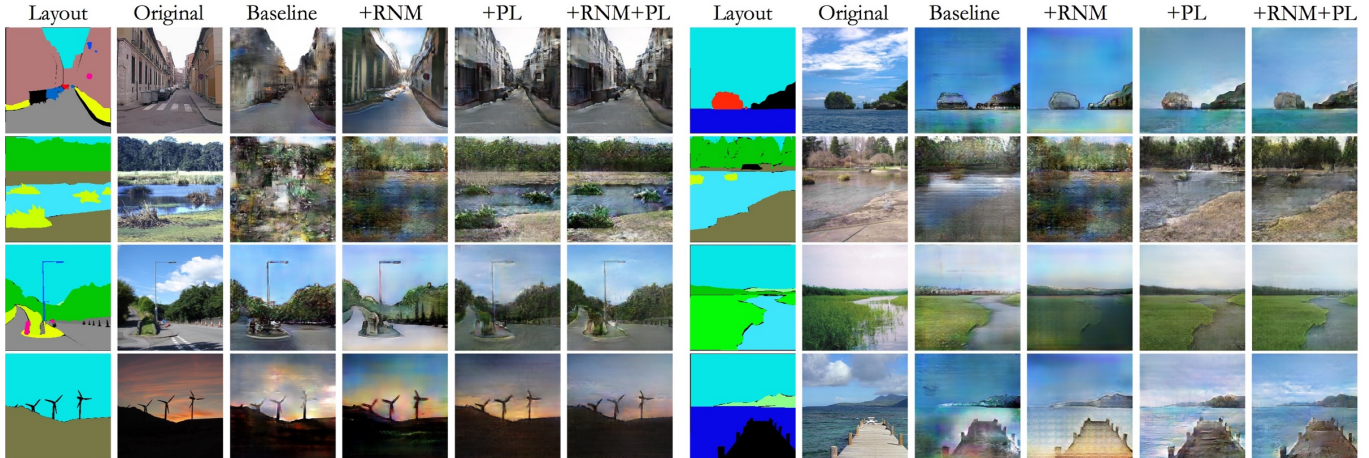
Fig. 4. Sample scene generation results. In these examples, the input layouts are from the test set, which are unseen during training and the transient attributes are fixed to the original transient attributes. Incorporating Relevant Negative Mining (RNM) and Perceptual Loss (PL) significantly improves the performance of the baseline SGN model in terms of both image quality as well as faithfulness of the end result to conditioned layouts and attributes.

For quantitative evaluation, we employ the Inception Score (IS) [Salimans et al. 2016] and the Fréchet Inception Distance (FID) [Heusel et al. 2017]. The IS correlates well with human judgment of image quality where higher IS indicates better quality. FID has been demonstrated to be more reliable than IS in terms of assessing the realism and variation of the generated samples. Lower FID value means that the distributions of generated images and real images are similar to each other. Table 1 shows the IS and FID values for our SGN model trained under various settings, together with values for the real image space. These results agree with our qualitative analysis that training with RNM and Perceptual Loss provides samples of the highest quality. Additionally, for each generated image, we also predict its attributes and semantic segmentation map using separately trained attribute predictor in [Baltenberger et al. 2016] and the semantic segmentation model in [Zhou et al. 2017] and we report the average MSE[2] and segmentation accuracy again in Table 1. Training with the proposed perceptual loss exhibits more advantages in preserving the desired attributes and the semantic layout.

Our SGN model with RNM and Perceptual Loss shows clear superiority to other variants both qualitatively and quantitatively. Thus from now on, if not mentioned otherwise, all of our results are obtained with this model.

4.1.3 Comparison with Image-to-Image Translation Models. We compare our model to the popular Pix2pix model of Isola et al. [2017]. We show qualitative comparisons in Figure 5. It is worth mentioning that Pix2pix generates images only by conditioning on the semantic layouts but not transient attributes, and moreover, it does not utilize the noise vector. For quantitative comparison, we compare the IS and FID scores and segmentation accuracy using all 1,338 testing images in Table 1. Furthermore, in addition to these metrics, we conduct a human evaluation on Figure Eight (formerly Crowdflower), asking workers to select among the results of our

Table 1. Quantitative results. We compare visual quality with respect to IS and FID, attribute and semantic layout correctness via average MSE of attribute predictions (Att. MSE) and segmentation accuracy (Seg. Acc.), respectively, via pre-trained models. Our SGN model trained with RNM and PL techniques consistently outperforms others, including the Pix2pix model [Isola et al. 2017].

Model	IS	FID	Att. MSE	Seg. Acc.
Pix2pix [Isola et al. 2017]	3.52	75.54	–	58.36
SGN	4.01	80.10	0.023	55.93
+RNM	4.22	84.28	0.028	51.90
+PL	4.04	60.70	0.016	**62.49**
+RNM+PL	**4.34**	**58.46**	**0.015**	62.39
Original	7.19	0	0.010	72.12

proposed model and the Pix2pix method (for the same semantic layout) which they believe is more realistic. 59% of the users picked our results as more realistic. These results suggest that besides the advantages of manipulation over transient attributes, our model also produces higher quality images than the Pix2pix model. We also tried comparing our results against the recently proposed Cascaded Refinement Network [Chen and Koltun 2017], however, it did not give meaningful results on our dataset containing complex scenes [3], hence we left these results out.

4.1.4 Diversity of the Generated Images. One can generate different version of the same scene by conditioning on different transient attribute vectors, while fixing the layout. Moreover, the proposed framework enables users to play with continuous attribute values, offering refined control over each specific attribute. In Figure 6, we show the effect of varying the transient attributes for a sample semantic layout. As can be seen, our model is capable of generating

---

[2]The ground truth attributes are scalar values between 0 and 1.

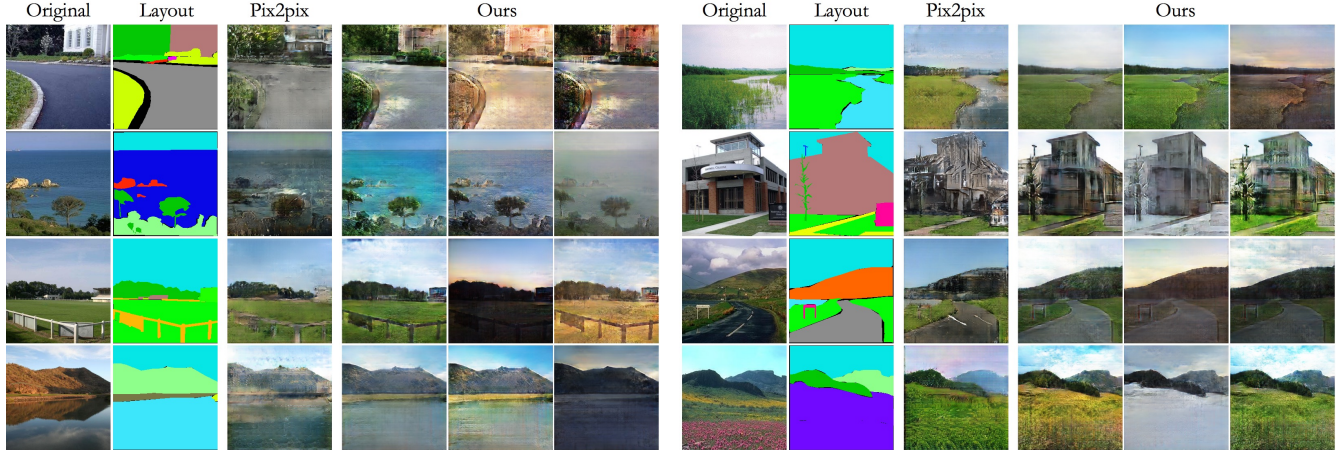[3]We trained this model using the official code provided by the authors,

Fig. 5. Comparison of our SGN model against Pix2pix [Isola et al. 2017]. In each row, we provide the original image which we use its semantic layout as input, the result of Pix2pix and three different results generated by our model with randomly chosen transient attributes. Our SGN model is able to generate diverse and more realistic results than Pix2pix.
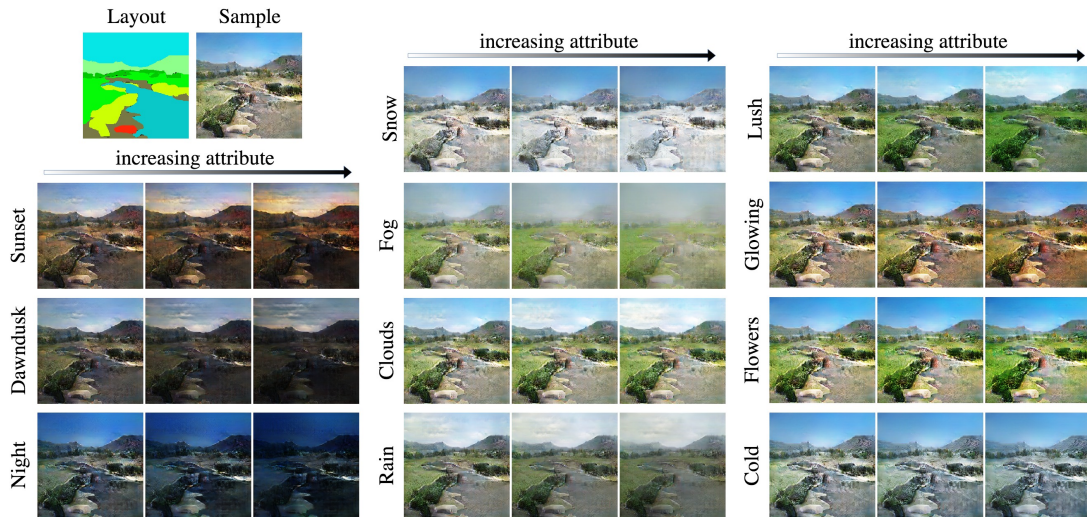

Fig. 6. Modifying transient attributes in generating outdoor images under different weather and time conditions. Our model's ability of varying with transient attributes contributes to the diversity and photorealism in its generation (more results can be found in the supplementary material).

diverse samples and it also enables us to manipulate the degree of desired transient attributes.

4.1.5 Adding and Subtracting Scene Elements. Here we envision a potential application of our model as an interactive scene editing tool that can add or subtract scene elements. Figure 7 demonstrates an example. We begin with a coarse spatial layout which contains two large segments denoting the "sky" and the "ground". We then gradually add new elements, namely "mountain", "tree", "water". At each step, our model inserts a new object based on the semantic layout. In fact, such a generation process closely resembles human thought process in imagining and painting novel scenes. The reverse process, subtracting elements piece by piece, can be achieved in a similar manner. Moreover, we sample different random attribute

vector to illustrate how generation diversity can enrich the outcomes of such photo-editing tools. We provide a video demo in the Supplementary Materials.

4.2 Attribute Transfer

We demonstrate some attribute manipulation results obtained with our approach in Figure 8. As can be seen, our algorithm produces photo-realistic manipulation results for many different types of attributes like "Sunset", "Spring", "Fog", "Snow", etc., and moreover, a distinctive property of our approach is that it can perform multimodal editing for a combination of transient attributes as well, such as "Winter and Clouds" and "Summer and Moist".
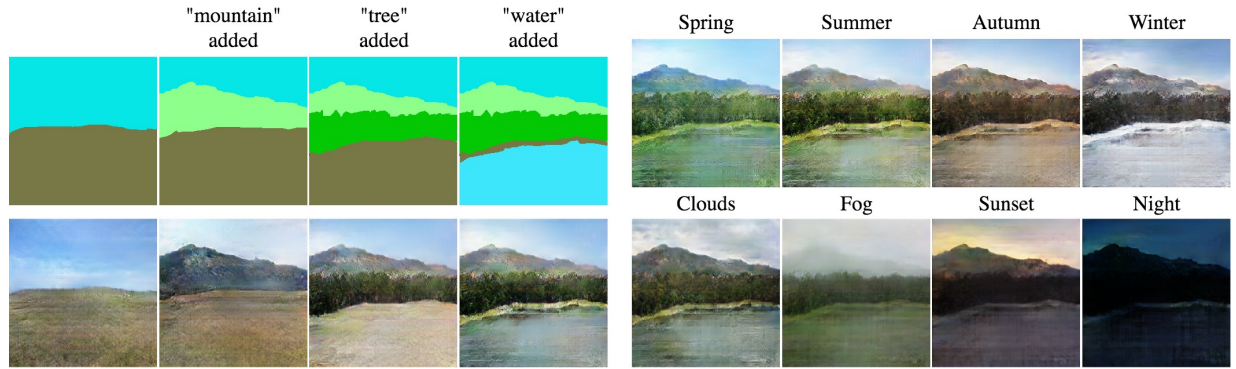
Fig. 7. Gradually adding and removing elements to and from the generated images. We use a coarse spatial layout map (top left) to generate an image from scratch, and then keep adding new scene elements to the map to refine the synthesized images. Moreover, we also show how we can modify the look by conditioning on different transient attributes.

Fig. 8. Sample attribute manipulation results. Given an outdoor scene and its semantic layout, our proposed framework is able to produce realistic looking results for modifying various different transient attributes. Moreover, our approach can perform multimodal editing as well, in which we modify a combination of attributes.

In Figure 9, we compare the performance of our method to the data-driven approach of Laffont et al. [2014]. As mentioned in Section 2, this approach first identifies a scene that is semantically similar to the input image using a database of images with attribute annotations, then it retrieves the version of that scene having the desired properties, and finally, the retrieved image is used as a reference for style transfer. Hence, in the figure, we both present the reference images generated by our approach and retrieved by the
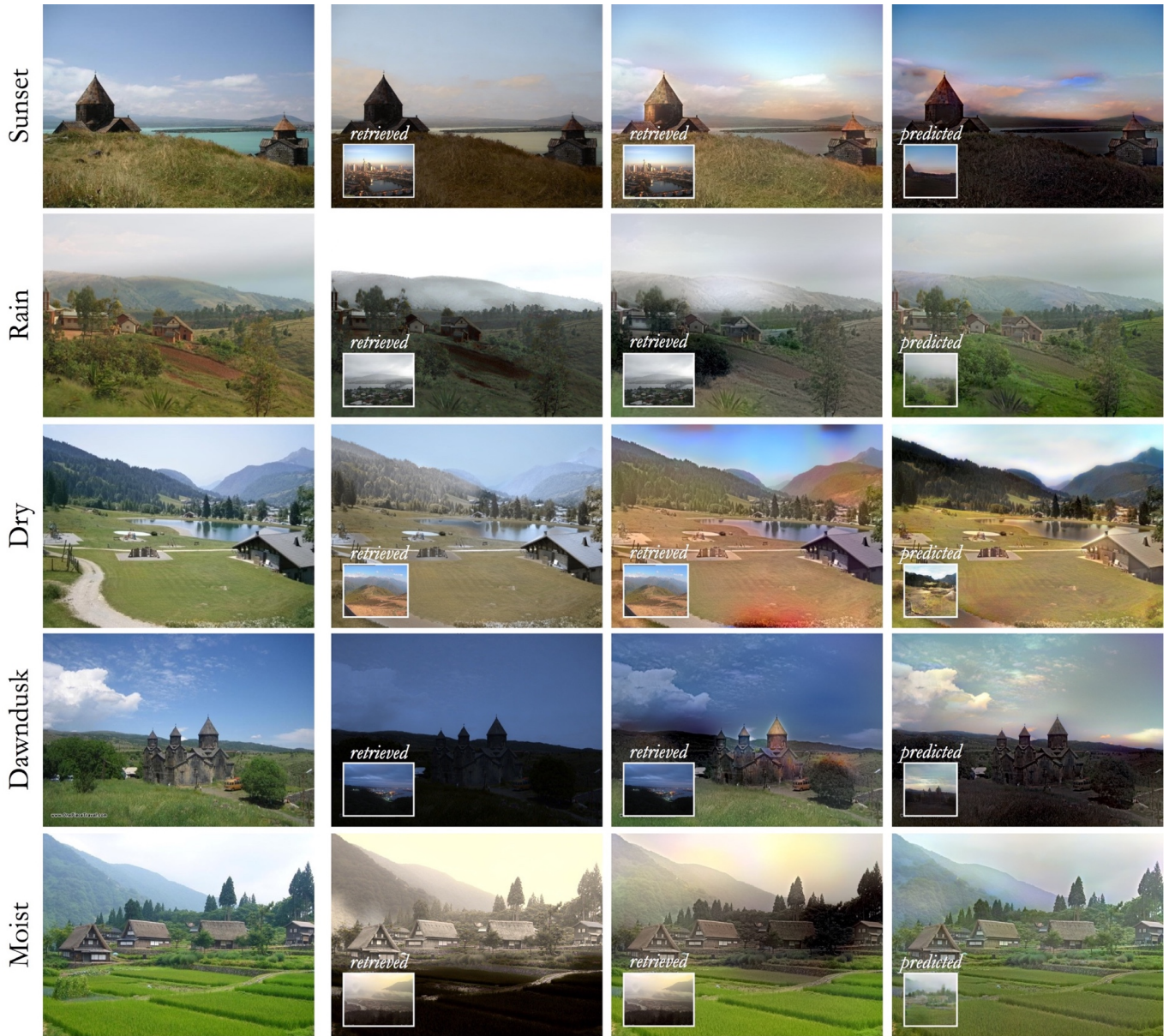
Fig. 9. Comparison with [Laffont et al. 2014]. For given input image, (a) and (b) : results of [Laffont et al. 2014] using the exemplar-based style transfer method they proposed and deep photo style transfer method [Luan et al. 2017] respectively between retrieved images and input images, c) : results of our method using deep style transfer method [Luan et al. 2017] between generated image by proposed SGN model and input image.

competing method at the right-bottom corner of each output image. For a fair comparison, we also present alternative results of [Laffont et al. 2014] where we replace the original exemplar-based transfer method with the deep photo style transfer method [Luan et al. 2017], which is used in obtaining our results[4]. As can be seen, our approach produces better results than [Laffont et al. 2014] in terms of visual quality and as to reflecting the desired transient attributes. These

results also demonstrate how style transfer methods are dependent on semantic similarity between the input and style images. In this regard, our approach provides a more principled way to edit an input image to modify its look under different conditions.

Additionally, we conducted a user study on Figure Eight to validate our observations. In this experiment, subjects are presented with results obtained with our approach and those of [Laffont et al. 2014] and are forced to select the one which they consider to be visually more appealing regarding the target attributes. We have a

[4]Note that, the post-processing method Mechrez et al. [2017] is also employed here to improve photo-realism.
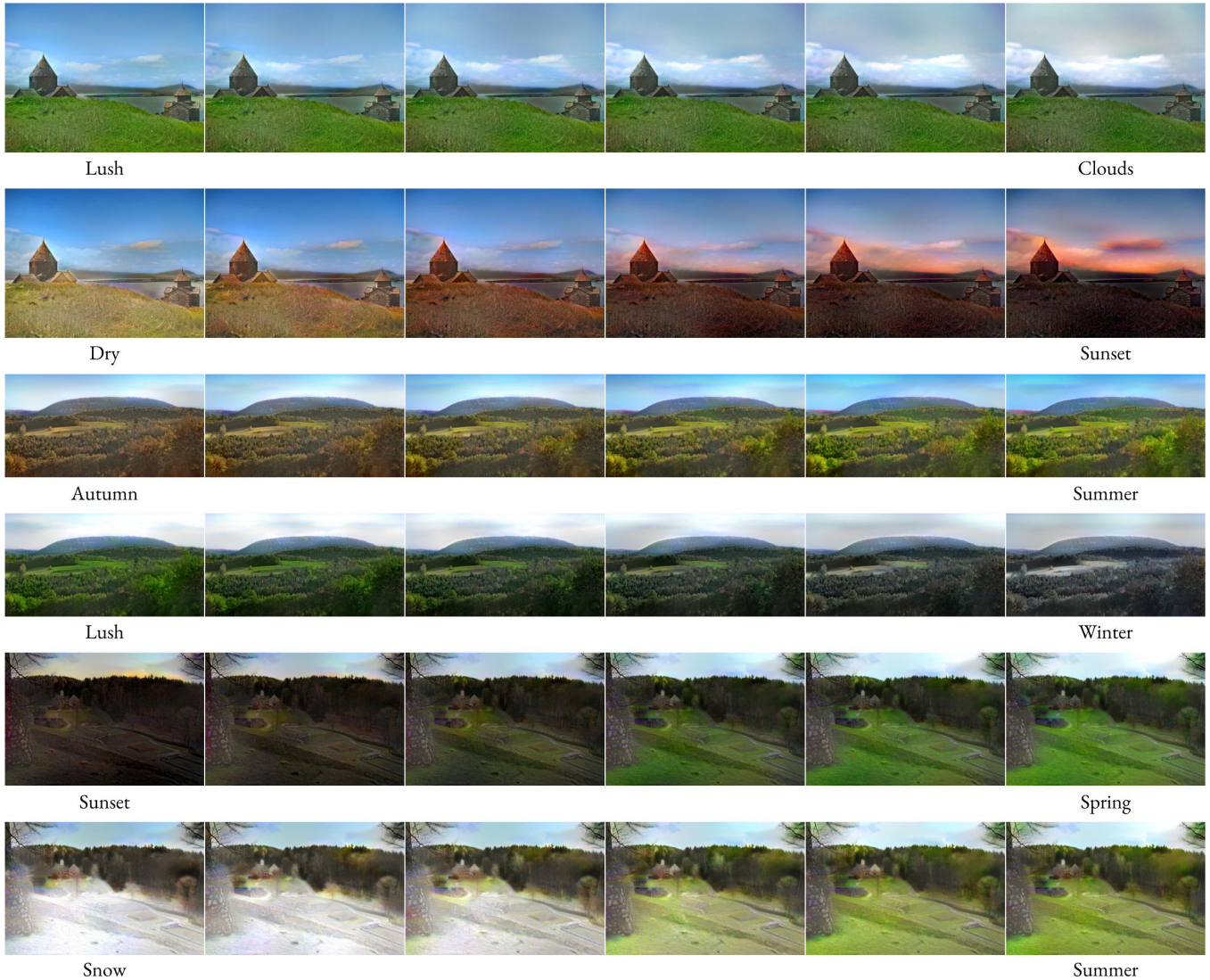
Fig. 10. Our method produces photo-realistic manipulation results for different degrees of transient attributes.

total of 48 questions and we collected at least 5 user responses per each of these question. Our results are favored 70% of the time (in 33 out of 48 comparisons).

The most important advantage of our framework over existing works is that our approach enables users to play with the degree of desired attributes via changing the numerical values of the attribute condition vector. As shown in Figure 10, we can increase and decrease the strength of specific attributes and smoothly walk along the learned attribute manifold using the outputs from the proposed SGN model. This is nearly impossible for a retrieval-based editing system since the style images are limited with the richness of the database.

It should be noted that modifying an attribute is inherently coupled with the emergence or disappearance of certain semantic content. For example, in urban scenes, increasing "night" attribute illuminates the buildings since lights are turned on at night, whereas increasing "sunset" attribute darkens the buildings; as another example, "clouds" attribute does not modify the global appearance of the scene but merely the sky region, comparing with "fog" attribute which blurs distant objects; "cold" attribute emphasizes the cold colors, while "warm" attribute has the opposite effect.

Consistent with the earlier discussion, our generated samples reflect the semantic meaning of the target attributes. To name a few, "sunset" attribute makes the horizon slightly more reddish, "autumn" attribute increases the brown tones on the trees, "snow" attribute whitens the ground. Also note that the emergence of each

Fig. 11. Season transfer to paintings. Source images: Wheat Field with Cypresse by Vincent van Gogh (1889), In the Auvergne by Jean-Francois Millet (1869) and Lourmarin by Paul-Camille Guigou (1868), respectively.
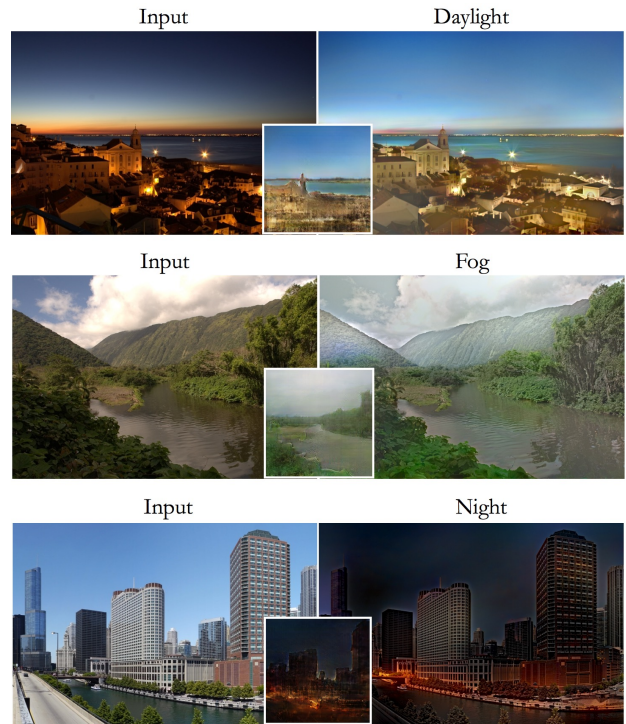

Fig. 12. Example failure cases for our attribute manipulation framework, which are due to failing of the photo style transfer method (top 2 rows) and the visual quality of synthesized reference style image (bottom row).

attribute tends to highly resonate with part of the image that is most related to the attribute. That is, "sunset" attribute primarily influences the sky, whereas "snow" attribute correlates with the ground, and "autumn" tends to impact the trees. This further highlights our model's reasoning capability about the attributes in producing realistic synthetic scenes.

Even though our attribute manipulation approach is designed for natural images, we additionally apply it to oil paintings in Figure 11. We change their attributes to obtain novel versions of these landscapes depicting different seasons. As can be seen, our model gives visually pleasing results, hallucinating what these paintings might look like if the painters painted the same scene at different times.

4.3 Limitations

Despite generally giving plausible results, our system does fail in some instances. Figure 12 demonstrates example failure cases which are due to shortcomings of the photo style transfer or the quality of the generated scene. In the first two, our SGN model hallucinated "daylight" and "fog" attributes successfully but the style transfer method fails to transfer the look to the input images. In the last, however, the generated scene does not reflect the "night" attribute perfectly, hence the corresponding manipulation result is not very convincing.

5 CONCLUSION

We have presented a high-level image manipulation framework to edit transient attributes of natural outdoor scenes. The main novelty of the paper is to utilize a scene generation network in order to synthesize on the fly the reference style image that is consistent with the semantic layout of the input image and exhibit the desired attributes. Trained on our richly annotated ALS17K dataset, the proposed generative network can hallucinate many different attributes reasonably well and even allows edits with multiple attributes in a unified manner. For future work, we plan to extend our model's functionality to perform local edits based on natural text queries, e.g. add or remove certain scene elements using referring expressions. For our sake, the hallucinated scenes do not need to be high-resolution, however our model can also benefit from the recent advances in image synthesis such as Progressive GANs [Karras et al. 2018].

ACKNOWLEDGMENTS

REFERENCES

Xiaobo An and Fabio Pellacini. 2010. User-Controllable Color Transfer. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 263–271.

G. Antipov, M. Baccouche, and J. Dugelay. 2017. Face Aging With Conditional Generative Adversarial Networks. In *The IEEE International Conference on Image Processing (ICIP)*.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

Ryan Baltenberger, Menghua Zhai, Connor Greenwell, Scott Workman, and Nathan Jacobs. 2016. A Fast Method for Estimating Transient Scene Attributes. In *Winter Conference on Application of Computer Vision (WACV)*.

David Berthelot, Tom Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

Floraine Berthouzoz, Wilmot Li, Mira Dontcheva, and Maneesh Agrawala. 2011. A Framework for content-adaptive photo manipulation macros: Application to face, landscape, and global manipulations. *ACM Transactions on Graphics(TOG)* 30, 5 (2011), 120–1.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature Verification using a "Siamese" Time Delay Neural Network. In *NIPS*.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. 2016. Hard negative mining for metric learning based zero-shot classification. In *European Conference on Computer Vision Workshops (ECCVW)*.

Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. 2014. ImageSpirit: Verbal guided image parsing. *ACM Transactions on Graphics (TOG)* 34, 1 (2014), 3.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kevin Dale, Micah K Johnson, Kalyan Sunkavalli, Wojciech Matusik, and Hanspeter Pfister. 2009. Image restoration using online photo collections. In *The IEEE International Conference on Computer Vision (ICCV)*.

Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems* (2013).

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *The IEEE Computer Vision and Pattern Recognition (CVPR)*.

Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 118.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems(NIPS)*.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. DRAW: A Recurrent Neural Network For Image Generation. In *International Conference on Machine Learning (ICML)*.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. 2016. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013* (2016).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems(NIPS)*.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. *arXiv preprint arXiv:1804.04732* (2018).

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 107.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*.

Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 94.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. 2014. Transient Attributes for High-Level Understanding and Editing of Outdoor Scenes. *ACM Transactions on Graphics (proceedings of SIGGRAPH)* 33, 4 (2014).

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2015. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300* (2015).

Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, and In So Kweon. 2016. Automatic content-aware color and tone stylization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xirong Li, CeesG M Snoek, Marcel Worring, Dennis Koelma, and Arnold WM Smeulders. 2013. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia* 15, 4 (2013), 933–945.

Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative face completion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 3.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 120.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep Photo Style Transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. 2016. Generating Images from Captions with Attention. In *International Conference on Learning Representations (ICLR)*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. 2016. Multi-class Generative Adversarial Networks with the L2 Loss Function. *arXiv preprint arXiv:1611.04076* (2016).

Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. 2017. Photorealistic Style Transfer with Screened Poisson Equation. In *The British Machine Vision Conference (BMVC)*.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and Checkerboard Artifacts. *Distill* (2016). https://doi.org/10.23915/distill.00003

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. *arXiv preprint arXiv:1606.05328* (2016).

Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. 2005. N-dimensional probability density function transfer and its application to color transfer. In *The IEEE International Conference on Computer Vision (ICCV)*.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.. In *International Conference on Learning Representations (ICLR)*.

Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016a. Learning What and Where to Draw. In *Advances in Neural Information Processing Systems (NIPS)*.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016b. Generative Adversarial Text to Image Synthesis. In *Internation Conference on Machine Learning (ICML)*.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. 2001. Color transfer between images. *IEEE Computer graphics and applications* 21, 5 (2001), 34–41.

Cynthia Rudin and Robert E Schapire. 2009. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research* 10, Oct (2009), 2193–2232.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. 2234–2242.

Ahmed Selim, Mohamed Elgharib, and Linda Doyle. 2016. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)* 35, 4 (2016), 129.

Wenling Shang, Kihyuk Sohn, Zeynep Akata, and Yuandong Tian. 2017. Channel-Recurrent Variational Autoencoders. *arXiv preprint arXiv:1706.03729* (2017).

Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. 2013. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 200.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating Videos with Scene Dynamics. In *Advances in Neural Information Processing Systems(NIPS)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2017. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. *arXiv preprint arXiv:1711.11585* (2017).

Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *The IEEE International Conference on Computer Vision (ICCV)*.

Fuzhang Wu, Weiming Dong, Yan Kong, Xing Mei, Jean-Claude Paul, and Xiaopeng Zhang. 2013. Content-Based Colour Transfer. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 190–203.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. 2016. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)* 35, 2 (2016), 11.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017a. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017b. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems (NIPS)*.