Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

USIS: Unsupervised Semantic Image Synthesis

George Eskandar[a,*], Mohamed Abdelsamad[a], Karim Armanious[a], Bin Yang[a]

[a] *Institute of Signal Processing and System Theory, University of Stuttgart, Stuttgart, Germany*

ARTICLE INFO

Article history:
Received August 19, 2022

Keywords: Generative Adversarial Networks, Semantic Image Synthesis, Unpaired I2I Translation

ABSTRACT

Semantic Image Synthesis (SIS) is a subclass of I2I (I2) translation where a photorealistic image is synthesized from a segmentation mask. SIS has mainly been addressed as a supervised problem. However, state-of-the-art methods depend on a massive amount of labeled data and cannot be applied in an unpaired setting. On the other hand, generic unpaired I2I frameworks underperform in comparison. In this work, we propose a new framework, Unsupervised Semantic Image Synthesis (USIS), as a first step toward closing the performance gap between paired and unpaired settings. We design a simple and effective learning scheme that combines the fragmented benefits of cycle losses and relationship preservation constraints. Then, we make the discovery that, contrary to I2I translation, discriminator design is crucial for label-to-image translation. To this end, we design a new discriminator with a wavelet-based encoder and a decoder to reconstruct the real images. The self-supervised reconstruction loss in the decoder prevents the encoder from overfitting on a few wavelet coefficients. We test our methodology on 3 challenging datasets and set a new standard for unpaired SIS. The generated images demonstrate significantly better diversity, quality and multimodality.

1. Introduction

Semantic image synthesis (SIS) is the task of generating high resolution images from user-specified semantic layouts. It is a recent application of Generative Adversarial Networks (GANs) that was introduced by Pix2PixHD[3] in 2017. In 2019, Spatially Adaptive Normalization or SPADE[4] was proposed as a better alternative generator design for the task and since then, the field has significantly grown. SIS opens the door to an extensive range of applications such as content creation and semantic manipulation by editing, adding, removing or changing the appearance of an object. By allowing concept artists and art directors to brainstorm their designs efficiently, it can play a pivotal role in graphics design. In addition, SIS can be used as a data augmentation tool for deep learning models, by generating training data conditioned on desired scenarios which might be hard to capture or reproduce in real-life (corner cases in autonomous driving like accidents).

In contrast to graphics engines and simulation tools, semantic image synthesis doesn't need either specialized training to use or intricate information like 3D geometry, materials or light transport simulation[5], because it learns directly from the collected real data. Moreover, a problem commonly seen in graphics engines is that the synthesized images look visually different from real data and models trained only on synthetic images do not generalize well due to the domain gap between the two data distributions[6]. That is why SIS can be of a significant advantage because it bypasses these problems. However, the task of semantic image synthesis has mostly been addressed as a supervised learning problem. Although state-of-the-art methods[2] can produce visually appealing high resolution images, they suffer from several drawbacks. Most importantly,

*Corresponding author:
e-mail: george.eskandar@iss.uni-stuttgart.de (George Eskandar)

Fig. 1: Our model USIS is a first step towards bridging the performance gap between paired and unpaired image-to-image translation in the context of SIS. CUT[1] is the state-of-the-art in unpaired GANs while OASIS[2] is the state-of-the-art in paired SIS.

they depend on a lot of annotated paired data which is expensive and time-consuming to acquire: the average annotation time for one frame in the Cityscapes dataset is 1 hour [7]. Furthermore, as labeled datasets are usually smaller than unlabeled datasets, supervised training restricts the generator's learned distribution to the distribution of real images which have labels. Unpaired training allows the usage of a larger number of real images, which exhibit a greater variation; thus, enhancing the generalization capability of the learned model.

Unpaired conditional GAN frameworks[8, 9, 10, 1, 11, 12] can be used for SIS, but they suffer from several drawbacks: (1) these approaches colorize semantic maps by color-coding each class (as opposed to one-hot encoding), which pushes the model to learn an artificial color-to-color mapping instead of a semantic-to-color mapping. (2) The unsupervised losses in the state-of-the-art force relationships between the labels and images that do not preserve the semantic content in the case of SIS. (3) The normalization layers in the architecture of unsupervised models wash away the semantic labels as noted in [4]. (4) Many unsupervised paradigms cannot use a one-hot layout representation for the labels because it requires a different encoder design, and they already use a shared encoder across domains. Consequently, the generated samples from these models suffer from poor quality. Another downside is the inability to generate photorealistic images when the number of classes in the dataset is too big.

In this work, we propose a framework named USIS, or Unsupervised Semantic Image Synthesis, which can synthesize multimodal realistic images from labels without the use of paired data (Figure 1). Multimodal generation is defined as one-to-many mapping or the ability to generate more than one image from a given label. To our knowledge, this is the first paper to address the problem of unpaired SIS on its own and not as an application to the more general I2I (I2) translation. The USIS framework can be trained on any two unpaired datasets generalizing the use of SIS. For instance, the labels from GTA-V dataset[13] could be used in training along with realistic images from Cityscapes[7] or KITTI[14]. This is not possible in the supervised setting, as the models trained on GTA-V labels can only produce GTA-V like images. By virtue of its design, the unpaired setting can help eliminate dataset biases and push the model towards a better multimodal generation. The generated samples along with their labels can thus be used as a data augmentation technique for semantic segmentation models. USIS also performs better on datasets with a large number of classes.

In the following, we first review the related works to our proposed framework. Then, after going through some preliminaries, a formal definition of the USIS task is given and some typical problems in the previous unpaired GANs are exposed and analyzed. Then, we present our framework USIS. In short, our contributions can be summarized as follows:

1. We address for the first time the problem of unpaired semantic image synthesis on its own, not as part of I2I translation.
2. We introduce a new unsupervised learning paradigm for GANs tailored to unpaired SIS. The paradigm involves an adversarial training between a generator and a whole image wavelet-based discriminator, and a cooperative training between the generator and a segmentation network (no pretraining). Although cycle-losses have been introduced before in CycleGAN, our approach is different because (1) it uses only a one-sided cycle loss and (2) it is not based on common L_1, instead it employs a class-balanced cross entropy loss.
3. In Section 5.1, we provide an explanation why this seemingly non-intuitive scheme works and is well-suited for unpaired label-to-image translation. In short, our approach is a simple and novel combination of two well-known paradigms: cycle losses and relationship preservation constraint.
4. We make the discovery that, contrary to unpaired I2I translation [8, 1], the discriminator design is crucial for model convergence and image quality.
5. To this end, we propose a novel discriminator design, that makes use of a well-known wavelet-encoder [15] and a newly designed wavelet decoder to reconstruct real images. The proposed decoder and reconstruction loss prevent the wavelet encoder from overfitting on a few wavelet coefficients in the image. Contrary to mainstream transposed convolutions in the decoder, we use wavelet upsampling layers instead. We make use of instance normalization layers instead of batch normalization. The decoder ends by an inverse wavelet transform layer and a tanh activation function.

6. Finally, extensive experiments on 3 image datasets (COCO-stuff[16], Cityscapes[7] and ADE20K[17]) in an unpaired setting are conducted to showcase the ability of our model to generate a high diversity of photorealistic images and close the gap between supervised and unsupervised methods in SIS. We set a new standard on all 3 benchmarks. We achieve more than two times the performance of the state-of-the-art on ADE20K and COCO-stuff. To the best of our knowledge, this is the first time an unpaired framework is able to generate reasonably photorealistic images on COCO-stuff and ADE20K from semantic labels, due to the massive amount of classes they contain ($\geq$ 150).

2. Related Works

Generative Adversarial Networks (GANs) GANs[18] can be trained to generate images. In GANs, two networks compete against each other in a minimax game. A generator tries to fool the discriminator into classifying the generated outputs as real. The past 5 years have witnessed great advances in the quality and resolution of the generated images. ProGAN [19] introduced the concept of progressive growing where GANs start with a few layers and are trained with low-resolution images. As the training continues more layers are added progressively to reach a higher resolution. StyleGAN[20] built upon ProGAN and fed style information in each layer of the generator to control visual features at different scales, using an adaptive instance normalization layer (AdaIN)[21]. StyleGANv2 [22] redesigned several aspects in the architecture like weight demodulation, path length regularization and progressive growing to improve upon StyleGAN. FastGAN [23] designed a light-weight generator for fast convergence along with a decoder on top of the discriminator. Similar to FastGAN, we deploy a decoder structure, which is however designed differently in order to match the wavelet nature of our discriminator. In all these models, the input to the generator is usually a random vector that follows a normal distribution and thus these models offer very little controllability in the generation process. Conditional GANs (cGANs) by contrast synthesize images based on a user-specified condition. Examples for conditions are class-labels[24, 25, 26], text[27, 28, 29] or other images[30, 3, 4, 8, 9, 31].

Supervised Semantic Image Synthesis is an image generation task where the condition is a semantic mask. The task was first introduced by Pix2pix[30]. The semantic map was color-coded and fed to an encoder-decoder architecture. A PatchGAN discriminator classifies overlapping patches in the generated images as real or fake. Chen et al. [32] proposed to use cascaded refinement networks and perceptual losses[33, 34] with a pretrained VGG network[35] for the task. Pix2PixHD[3] further improved the quality of generated results by employing feature matching losses to stabilize GAN training, a multiscale discriminator and a more sophisticated generator architecture. But the breakthrough came with SPADE[4] which realized the inadequacy of using normalization layers with semantic labels. To remedy the issue, Park et al.[4] proposed to generate the modulation parameters of the normalization layers from the semantic layout in a pixelwise manner. Moreover, the parameters vary spatially. Other choices have also been proposed to remedy the issue. For instance, CLADE[36] proposed to use class-adaptive modulation parameters instead. CC-FPSE[37] employed spatially-varying convolutional weights instead of the spatially-varying normalization layers. SEAN[38] used the SPADE layer but redesigned the network to add more controllability to edit the style of each semantic region individually. Similar to advances in the generator architecture, various improvements in the discriminator architecture have been proposed, even though perceptual losses were a standard in all these frameworks. OASIS[2] were the first to utilize a UNet-based discriminator [39] to improve the semantic image synthesis task. This "segmentation" discriminator, previously used to improve semantic segmentation[40] or unconditional image generation[41], tries to classify each pixel of real images into its semantic class and generated images as fake.

Unpaired I2I translations (U-I2I) is a conditional image generation task where it is either impossible or expensive to collect paired data. There has been two main approaches to solve this problem: using a cycle consistency loss or imposing a relationship preservation constraint. Cycle consistency aims to find correspondences between the input and output, by learning the inverse mapping and reconstructing the input[8, 42]. The CycleGAN[8] framework first introduced this approach and consisted of two generators (forward and inverse mapping) and two discriminators (one for each dataset). Research in this area has leveraged cycle consistency losses to allow for many-to-many mappings[43], multimodal mapping between two domains[10, 44, 9, 45] and an improved generation quality[46, 47, 48, 49, 50]. In many of these frameworks, the image data was assumed to be generated from a content and a style latent variables. MUNIT[9] mapped an image from domain A to domain B by combining its content code and a sampled style code from domain B. Cycle losses were applied not on images but rather on the latent codes. However, the problem with cycle losses, is that they assume that a mapping from one domain to another is a bijection which is often a restrictive assumption. On the other hand, some works[51, 52, 53, 12, 54, 11] have approached this task by imposing a relationship preservation constraint. If x_1 and x_2 are 2 images in domain A with a certain relationship $\mathcal{R}$, their mappings in domain B, $G(x_1)$ and $G(x_2)$, should have the same relationship. This constraint doesn't have to happen only on an image level, it can also occur between patches of the same image[50, 1]. In some works, the relationship constraint is a predefined distance loss (content losses[51, 52, 53] or geometric constraints[11]), in others it is based on a contrastive loss [54, 1]. Our work can be considered as a one-sided cycle-consistency loss and a relationship preservation constraint on a pixel-level at the same time.

Frequency-based Approaches in Deep Learning have gained more attention in recent years. The bias of convolutional neural networks (CNNs) towards low-frequency has been studied in several works[55, 56, 57]. More specifically, in [55] it has been observed that the discriminator is missing high frequency

information due to the downsampling operations in the network architecture while in [56], it was found out that upsampling layers in the generator cannot reconstruct the spectral distribution of the real data. As a remedy to these issues, high frequency representations of images like 2D-Fourier transform[58] and Haar wavelet transform[59, 60] have been used with neural networks either as components in the architecture in several generative applications like image super-resolution, image denoising or style-transfer[61, 62, 63, 64, 65, 66]. Other works have based their generative networks on high-frequency representation. For instance, [67] proposed a discriminator that observes the wavelet decomposition of generated and real images while [68] proposed a generator that operates in the wavelet domain. Some works used the wavelet representation in both the generator and discriminator. MW-GAN[69] proposed a multi-level wavelet generator and a discriminator that evaluates the images in the spatial and wavelet domains. WaveletSRGAN[70] presented a wavelet generator, a pixel-level discriminator and a wavelet-level discriminator for super-resolution task. SWAGAN[15] was the first unconditional GAN to incorporate wavelet operations in the generator and discriminator of StyleGANv2[22] and has shown promising results. We notice that the wavelet representation has been consistently preferred over the Fourier representation in deep learning models, due to the fact that the wavelet decomposition offers simultaneous information in both the space and spectral domains making it more suitable to CNNs. Wavelet decomposition also offers a multi-resolution analysis and has a faster computation time than the Fourier transform. However, none of the previously mentioned works used wavelet representations in an unsupervised setting: the wavelet decomposition of the groundtruth is always available for reconstruction. This is in contrast to our framework because our input consists of label maps whose wavelet transform has little or no semantic value.

3. Preliminaries

To make the paper self-contained, we briefly review in this section some fundamental concepts about semantic image synthesis and wavelet transform.

3.1. The SPADE Baseline

SPADE [4] is a GAN which consists of two components: (1) a generator with a decoder-like structure which cascades several ResNet blocks[71] with upsampling layers in between, and (2) a multiscale Patch-discriminator. The input to the generator is a random vector sampled from a multivariate Gaussian distribution while the semantic map is fed to the SPADE layer in each ResBlock after being downsampled to the corresponding resolution. Normalization layers are replaced by the SPatially Adaptive DEnormalization layers (SPADE), where the features f coming from convolutional layers are first normalized per channel and modulated with spatially variant learned parameters from semantic maps. More concretely, the semantic maps pass through a few convolutional layers to produce two tensors: γ and β. After normalization, the output o of SPADE is:

$$
\begin{aligned}
o_{b,c,i,j} &= \gamma_{c,i,j}(\mathbf{m})\frac{f_{b,c,i,j} - \mu_c}{\sigma_c} + \beta_{c,i,j}(\mathbf{m}) \\
\mu_c &= \frac{1}{BHW}\sum_{b,i,j} f_{b,c,i,j} \\
\sigma_c &= \sqrt{\frac{1}{BHW}\sum_{b,i,j}(f_{b,c,i,j} - \mu_c)^2}
\end{aligned}
\tag{1}
$$

where $\mathbf{m}$ is the semantic layout, B, H and W are the batch-size, height, and width of the features, respectively, and $b \in B, c \in C, i \in H, j \in W$. The losses of SPADE are similar to Pix2PixHD[3] and consist of a GAN loss and a feature matching loss[72, 34, 33] for each discriminator, and a perceptual loss based on a pretrained VGG-network on ImageNet[73].

3.2. The wavelet transform

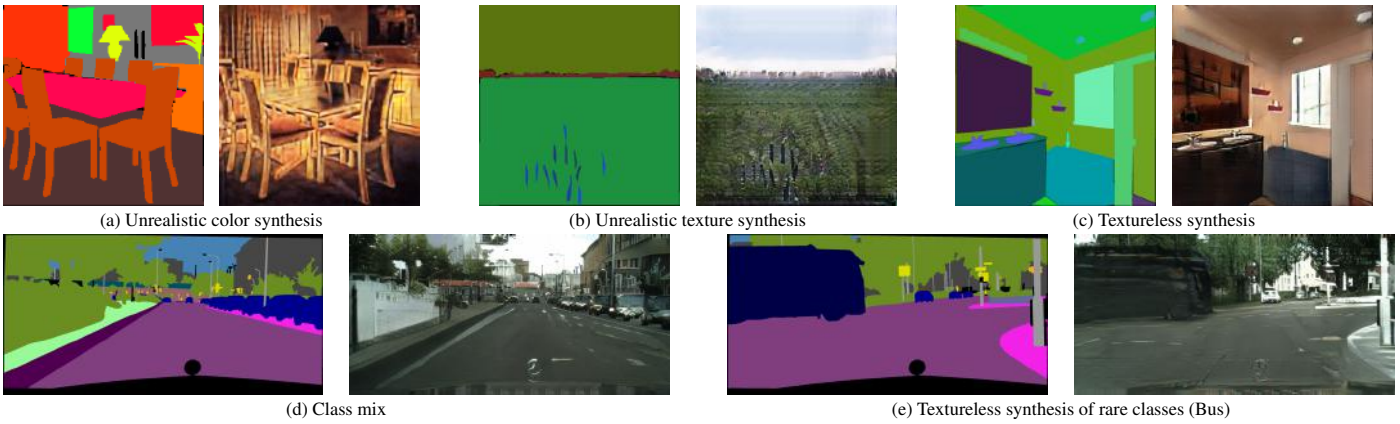
Our work is strongly based on the wavelet transform, which passes the image through a series of low-pass and high-pass filters to generate Haar wavelet coefficients. The Haar wavelet is a family of functions that form an orthonormal basis which can represent discrete signals. One-level wavelet decomposition generates 4 subbands of lower resolution: an LL frequency subband, which is a blurred version of the image, and 3 high frequencies subbands: LH,HL and HH which represent higher frequencies in horizontal, vertical and diagonal directions, respectively. The wavelet transform can be applied recursively to the LL subband to generate 4 more subbands at a smaller resolution, thus offering a multi-frequency multi-resolution analysis of the image. In deep learning, there has been 2 main ways to exploit wavelets: either by employing a multi-level decomposition and a reconstruction loss[70], or by using only one-level decomposition at a time and progressively growing the network like in [15].

In this work, we seek to synthesize an image from a semantic map in an unsupervised way. Since there is no direct feedback signal from the groundtruth, the network can favour the generation of big classes (streets, buildings) over small ones (pedestrians, traffic signs) to minimize the GAN objective. Providing the network with a high-frequency representation is advantageous for the unpaired setting because small objects in the image domain have coefficients with bigger magnitude in the high-frequency subbands of the wavelet transform. By exploiting this property in the proposed framework we seek to accomplish two objectives: (1) to generate more fine-details and refine the texture of bigger semantic classes and (2) to foster the generation of smaller semantic classes.

4. Problem Definition

In this section, a formal definition of the SIS task is presented along with some common problems found in the previous unpaired baselines; all of which lays the groundwork for the proposed model in the next section.

In SIS, we seek to synthesize an RGB image $\mathbf{x}$ from a semantic mask $\mathbf{m}$ with C labels in an unsupervised way. Let i denote



(a) Unrealistic color synthesis (b) Unrealistic texture synthesis (c) Textureless synthesis

(d) Class mix (e) Textureless synthesis of rare classes (Bus)

Fig. 2: Common challenges in SIS. The shown images were generated from CUT[1] and CycleGAN[8] on Cityscapes and ADE20K. The cycle consistency penalty in CycleGAN leads to a faulty semantic correspondence in Cityscapes (d) and an unrealistic color generation in ADE20K (a). CUT suffers from either a textureless synthesis (c) or an unrealistic texture synthesis (b). (e) is a common problem in the supervisd and unsupervised settings.

the pixel position. The images are normalized between 0 and 1, and $\mathbf{m}$ is one-hot encoded. The problem can be broken down to two tasks:

- Class appearance matching: how will the discriminator know the correspondence between the semantic class in the segmentation map and its appearance (texture) in the image without supervision? and how to learn this mapping in a multimodal way?
- Semantic alignment: how can the generator preserve the content/geometrical structure of the segmentation map without ignoring small classes ?

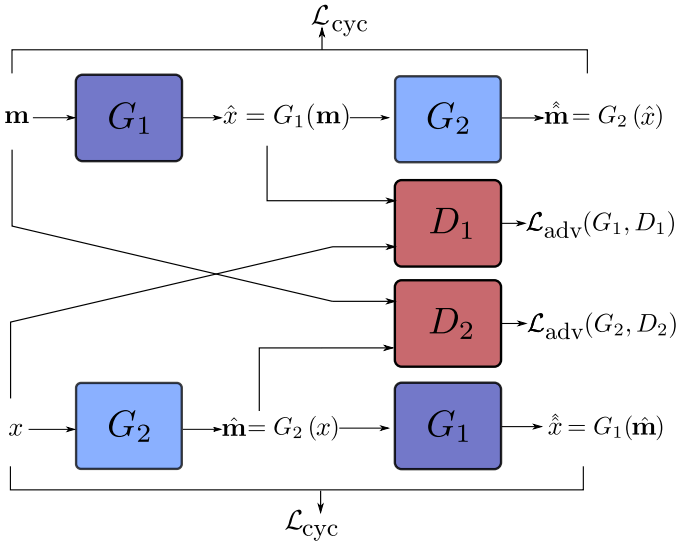
The reason why unpaired GANs are suboptimal is that they color-code the segmentation map $\mathbf{m}$ and feed to the network as an RGB image with values between 0 and 1. This way they learn a mapping between color information ($\mathbf{m}_{i,j} \in [0,1]^3 \longrightarrow \mathbf{x}_{i,j} \in [0,1]^3$) instead of learning a mapping between a semantic label and color information ($\mathbf{m}_{i,j} \in \{0,1\}^C \longrightarrow \mathbf{x}_{i,j} \in [0,1]^3$). One might think that it is possible to use a one-hot encoded representation with these frameworks; however this is not possible in many of the current unpaired GANs [10, 1, 9, 12] because of their architectural design. Other frameworks [8, 11] can be modified to use a one-hot encoded representation, but we show in Section 6.2 that these learning schemes are suboptimal for this task.

The two main paradigms of unpaired GANs presented in Section 2, cycle-consistency and relationship preservation, suffer from a few issues when applied to color-coded labels. Traditional cycle-consistency Mean Absolute Error (MAE) or Mean Squared Error (MSE) losses can preserve alignment between the segmentation map and the image but might lose semantic information (for example, buildings are generated instead of trees or sky). Furthermore, in SIS, the reverse cycle (RGB to semantic map to RGB) is redundant because it might not always produce segmentation maps, but rather copies the same RGB image with the texture or style of semantic layouts (as observed in CycleGAN results). On the other hand, relationship preservation constraints may result in more diverse images (reflected in the FID score) but still suffer from the same problem (good spatial alignment with loss of semantic information). The relationship preservation is usually a constraint imposed between the input label and the generated image in the form of a predefined distance[12] or a contrastive loss on the features of an encoder network[1]. For instance, CUT[1] maximizes the mutual information between features extracted from corresponding patches in input and output. However, it has been shown[74] that CNNs are biased toward the texture of the image rather than the shape. This way, we see that color information originating from the color-coding of the classes gets leaked into the features of the encoder and affects the contrastive loss. Finally, a common problem in both approaches is that the normalization layers wash away the semantic information [4] and convolutional layers are biased towards low frequencies.

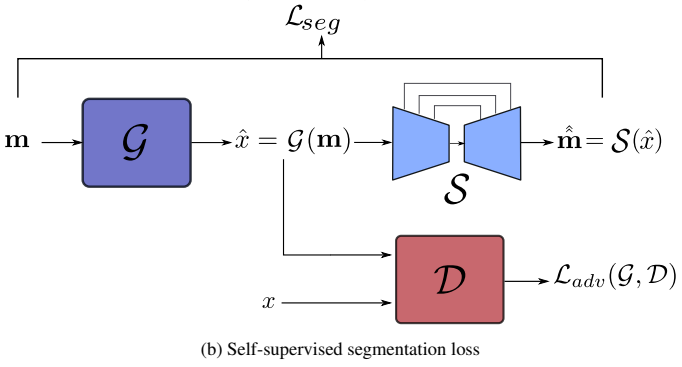
The most common problems that occur in unsupervised semantic image synthesis can be summarized in the following list and visualized in Figure 2:

- Class Mixing: class A appears instead of class B
- Unrealistic color synthesis
- Noisy texture synthesis
- Textureless objects (problems identified and solved by SPADE[4] in the supervised setting)
- Loss of fine details in the object or unrealistic appearance: this problem also occurs in the supervised setting and is related to the low-frequency nature of convolutional layers.
- Textureless synthesis for rare classes in the dataset

The first 2 problems occur more often in the unsupervised setting, because the discriminator doesn't have a direct feedback that allows for class recognition. The rest is common in both supervised and unsupervised settings. We assume that the greater the number of classes is in a dataset, the more these problems appear because the classes become harder to distinguish when they have more similar color codes. Moreover, the number of rare classes increases in a large and diverse dataset. This is shown in section 6.



(a) Cycle losses in Cycle-GAN[75]



(b) Self-supervised segmentation loss

Fig. 3: Cycle losses and self-supervised segmentation loss in USIS. A generated image should appear realistic to the discriminator $\mathcal{D}$ while generating classes that could be distinguished by $\mathcal{S}$.

5. Proposed Framework

In this section, we propose a novel framework, USIS, for unpaired semantic image synthesis, which builds upon the fragmented benefits of the cycle-consistency approach and the relationship preservation approach. We first introduce the proposed unsupervised paradigm. Then, we explain how the self-supervised segmentation loss helps preserve the semantic alignment and enhance the ability of the generator to match the appearance of real data. Finally, we analyze how the design of the discriminator influences the unsupervised learning.

5.1. USIS Paradigm

The proposed framework consists of a UNet segmentor which is trained cooperatively with the generator by the means of a self-supervised segmentation loss; and a whole image wavelet-based discriminator which is trained adversarially to capture the color and texture distribution of all semantic classes.

The three components of our framework are: (1) a wavelet SPADE Generator $\mathcal{G}$, (2) a wavelet Discriminator $\mathcal{D}$ and (3) a UNet segmentation network $\mathcal{S}$. The generator generates an RGB image $\mathbf{x}$ from the semantic map $\mathbf{m}$ (one-hot encoded), the discriminator makes a decision whether the generated image is real or fake while $\mathcal{S}$ tries to segment the generated image back to the mask $\mathbf{m}$. We note that $\mathcal{S}$ only observes the generated images unlike the discriminator which sees real and generated images. Thus $\mathcal{D}$ competes with $\mathcal{G}$ to encourage the generation of photorealistic images while $\mathcal{S}$ and $\mathcal{G}$ cooperate to achieve semantic alignment in the form of a class-balanced [2] self-supervised segmentation loss, $\mathcal{L}_{seg}$, with the input mask. The loss for each component in our framework can be expressed as:

$$\begin{aligned}\mathcal{L}_{Gen} &= \lambda_{seg}\mathcal{L}_{seg}(\mathbf{m}, \mathcal{S}(\mathcal{G}(\mathbf{m}))) + \mathcal{L}_{adv_G}(\mathcal{D}(\mathcal{G}(\mathbf{m})))\\ \mathcal{L}_{UNet} &= \mathcal{L}_{seg}(\mathbf{m}, \mathcal{S}(\mathcal{G}(\mathbf{m})))\\ \mathcal{L}_{Dis} &= \mathcal{L}_{adv_D}(\mathcal{D}(\mathbf{x}), \mathcal{D}(\mathcal{G}(\mathbf{m})))\end{aligned} \tag{2}$$

where $\mathcal{L}_{seg}$ is expressed as:

$$-\mathbb{E}_{\mathbf{m}}\left[\sum_{c=1}^{C}\alpha_c\sum_{i,j}^{H\times W}\mathbf{m}_{c,i,j}\log(\mathcal{S}(\mathcal{G}(\mathbf{m}))_{c,i,j})\right] \tag{3}$$

The class-balancing weights α_c are proportional to the inverse of the per-pixel class-frequency.

$$\alpha_c = \frac{H\times W}{\sum_{i,j}^{H\times W}\mathbb{E}_{\mathbf{m}}\left[\mathbb{1}[\mathbf{m}_{c,i,j}]\right]} \tag{4}$$

The class balancing makes sure that smaller-classes have a strong contribution in comparison to bigger classes in the segmentation categorical cross entropy loss.

We adopt the non-saturating version[18] of the GAN logistic loss $\mathcal{L}_{adv}$, where the discriminator tries to maximize the probability of classifying the images $\mathbf{x}$ as real and the generated images $\mathcal{G}(\mathbf{m})$ as fake; and the generator tries to maximize the probability that the discriminator classifies $\mathcal{G}(\mathbf{m})$ as real.

$$\begin{aligned}\mathcal{L}_{adv_D} &= -\mathbb{E}_{\mathbf{x}}\left[\log(\mathcal{D}(\mathbf{x}))\right] - \mathbb{E}_{\mathbf{m}}\left[\log(1-\mathcal{D}(\mathcal{G}(\mathbf{m})))\right]\\ \mathcal{L}_{adv_G} &= -\mathbb{E}_{\mathbf{m}}\left[\log(\mathcal{D}(\mathcal{G}(\mathbf{m})))\right]\end{aligned} \tag{5}$$

We adopt the same regularization scheme as in StyleGANv2[22] to stabilize the training. An $\mathcal{R}1$ regularization[76, 24, 77] is applied to the discriminator every 16 minibatches. We design the discriminator to output a single score for the whole image, which stands in contrast to other conditional GANs[3, 8, 11, 12, 1, 4] that use patch discriminators.

Note that the self-supervised segmentation loss is different than the cycle losses in CycleGAN[8], MUNIT[9], and DRIT[10]. We do not seek to generate segmentation maps or do the inverse mapping. There is neither a discriminator for semantic maps nor a reverse cycle (Real image $\longrightarrow$ Segmentation map $\longrightarrow$ Real image). The two paradigms are depicted in Figure 3.

Why does USIS learning paradigm work ? One-sided cycle losses have received little attention, possibly because at first glance it is not obvious that they work, and because it has been shown in CycleGANs that they are suboptimal [8]. However, label-to-image translation is different from I2I translation as there exists much less information in the input label than the output image. The proposed paradigm addresses this fact in two ways.

First, the self-supervised segmentation loss heavily punishes the inseparability between regions belonging to different semantic labels and prevents the class-mixing problem. In the beginning of the training, it is easier for the generator to produce realistic images by matching the appearance of big classes while ignoring small ones. The self-supervised loss pushes the generator to synthesize small classes and to achieve a better semantic alignment to counteract the tendency of the generator to satisfy the GAN objective by finding a trivial solution (like matching one or two big classes to make the image appear realistic).

The second motivation is that the self-supervised segmentation loss can also be seen as a relationship preservation constraint. In previous works on unpaired GANs, the relationship was either defined between different images of one domain[54, 12] , transformations of the same image[11] or patches of the same image[1]. Instead of contrasting different output patches against each other like CUT[1], we contrast different pixels against each other with the help of the self-supervised segmentation loss. Our assumption is that pixels belonging to the same class-label should have similar features in the generator while pixels belonging to different class-labels should have dissimilar features. This is encouraged by classifying the generated pixels back to their labels. The preservation of a semantic relationship between different pixels of the image improves the generation capability of the network. In a previous work by Collins[78], a spherical k-means clustering has been conducted on the deep features of unconditional generative models like Progressive GANs and StyleGAN and has revealed that the feature clusters of a good generative model spatially span semantic objects. The proposed unsupervised paradigm encourages feature clustering explicitly: intra-class feature similarity is enabled by the SPADE layers, while the self-supervised loss enforces the inter-class feature separability inside the generator to ensure a higher generation quality and diversity.

5.2. *Discriminator Design*

The discriminator is an essential part of the framework because it is responsible for capturing the data statistics. Most importantly, it prevents the generator from learning trivial mappings (like identity mapping) that minimize the self-supervised segmentation loss; and it is the part responsible for discovering the appearance and texture of different classes in the scene in an unsupervised way.

An important design feature in the discriminator is its visual receptive field. Previous unsupervised models were mostly dependent on patch discriminators, which classify patches of size $N \times N$ pixels in the original image. The motivation for this design choice was to model the image as a Markov random field assuming that pixels separated by more than a patch size are independent. PatchGANs would thus capture high frequency content in the image, like its texture. However, we argue that the PatchGAN paradigm is not optimal for the purpose of unsupervised SIS because the discriminator is incapable of sufficiently penalizing individual confined objects with unrealistic texture when it has only a localized view of the image. This is due to its intrinsic design of averaging out the scores of all individual image patches which dilutes its sensitivity to local errors. As a result, PatchGANs in the context of SIS tend to match only the color distribution of real images while neglecting the texture distribution.

To counteract the drawbacks of PatchGANs, we propose the utilization of whole image discrimination. The whole discriminator assigns a bigger penalty to images with small unrealistic objects even if the remainder of the image has photorealistic textures. This situation happens particularly in the beginning of the training. However, when the training progresses, the whole image discrimination keeps providing the generator with a strong feedback signal to keep generating finer and smaller classes (humans, poles, traffic signs, ...).

Wavelet discrimination has been a valuable tool in supervised conditional image generation[68, 69, 70] and unconditional image synthesis[15]. In this work, we illustrate its usefulness in the unsupervised setting. The motivation for a frequency-based discrimination is that class appearance is multimodal and objects belonging to the same class may vary a lot across the dataset making it hard for the discriminator in an unsupervised setting to capture all variations of an object in regard to its scale, style, texture, pose and illumination. The smaller and finer an object is, the harder it is to render it in a photorealistic way without a direct supervision signal. To further enhance the capability of the discriminator network in rendering photorealistic small objects, the whole image discriminator architecture is also extended with the use of wavelet-based representations. More specifically, we incorporate the SWAGAN discriminator architecture, which was previously proposed in [15] to enhance the texture of generated images. Notably, we repurpose this architectural design to be used for unsupervised class appearance matching because it is more suited for discriminating high-frequency content while focusing on the whole image, in contrast to PatchGANs. By allowing the discriminator to process the discrete wavelet transform (DWT) of the image, the higher frequencies are not entirely lost in the downsampling layers of the discriminator and consequently the smaller classes can now have a bigger contribution in the adversarial loss function.

The SWAGAN discriminator[15] differs from other previously proposed wavelet discriminators[67, 69, 70]. It doesn't just use an n-level wavelet decomposition as input to the network, but rather downsamples the image multiple times in the pixel domain, performs a DWT on each resolution then maps it to high-dimensional features using a convolutional block (fWavelet). As can be seen in Figure 4, features from different resolutions are mapped together using skip connections. This architecture offers two advantages: first, it performs all n-level wavelet decomposition, one at a time in the network aggregating multiscale features instead of just performing 3- or 4- level wavelet decompositions at the input of the network. Second, this architecture was inspired from StyleGANv2 and designed to replace progressive growing[19] while retaining its advantage: to initially focus on low-resolution features then produce sharper details. While we use the discriminator without modifying its architecture, its motivation and role are different from SWAGAN: in unconditional image generation, objects of dif-

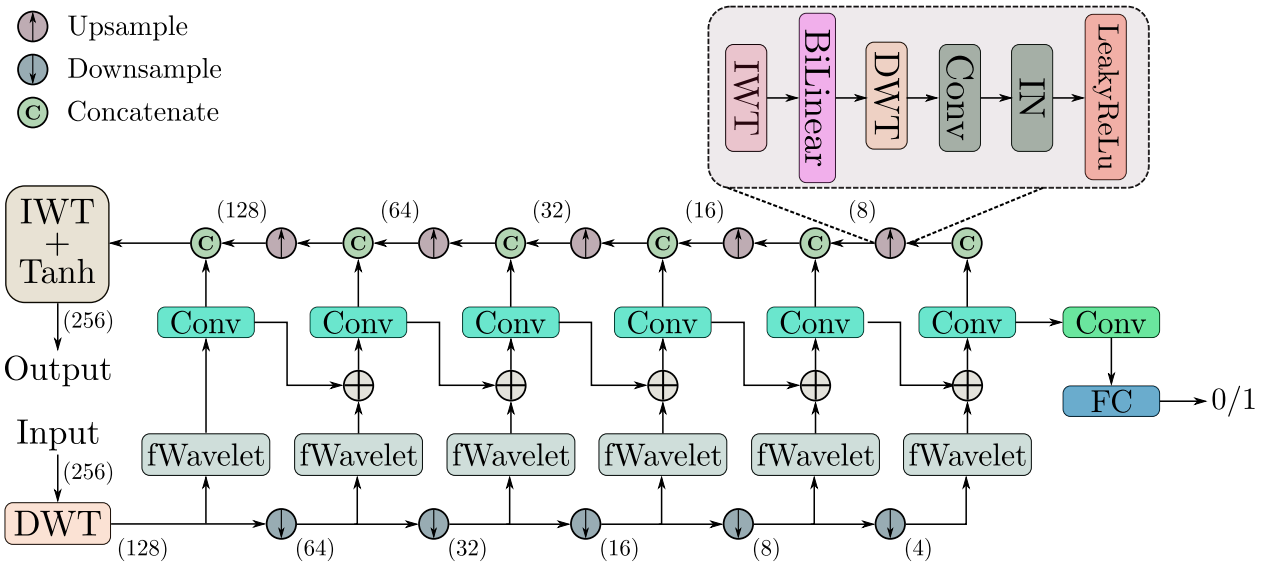


Fig. 4: Wavelet discriminator with decoder (upper branch) architecture. fWavelet denotes "from Wavelet" and refers to normal convolution layers that transform the image from wavelet domain to a higher dimensional feature space, in order to be processed further by the network. Bilinear refers to bilinear interpolation of the image. Bilinear interpolation is used for downsampling in the lower part of the network, and is used for upsampling in the upper part (decoder). IWT and DWT refer to inverse wavelet transform and discrete wavelet transform. FC is a fully connected layer. The numbers between brackets refer to the height dimension of the images/features.

ferent scales can be generated but suffer from a loss of high-frequency details. In our case, the generator in the unsupervised setting cannot even synthesize some of the classes in the dataset without the wavelet discriminator. Typically, these are classes that occupy a small scale.

To encourage the discriminator to consider the entire wavelet decomposition of the image, we attach a decoder on top of the discriminator. The decoder learns to reconstruct the real images only with an L_1-loss, and is updated during the discriminator gradient step. By learning to reconstruct the image like an autoencoder, the discriminator doesn't overfit as strongly on few wavelet coefficients. However, a straight-forward decoder design featuring transposed convolutions and batch normalization layers doesn't lead to any improvements. Since the discriminator operates in the wavelet domain, we redesign the upsampling layers (see Figure 4). The newly designed WaveletUpsample blocks (WU) consist of a cascade of IWT, bilinear upsampling in the image space, DWT and conv layer. This design choice is motivated by the fact that transposed convolution might upsample each wavelet subband differently, although they represent the same spatial information. In contrast, WU blocks maintain a spatial consistency in the upsampling. Lastly, we notice that using batch normalization layers make the decoder sensitive to the batchsize. To remedy this effect, we replace them by instance normalization layers.

6. Experiments and Discussion

We conduct our experiments on 3 datasets: Cityscapes[7], COCO-stuff[16] and ADE20K[17]. Cityscapes contains street scenes in German cities with pixel-level annotations of 19 classes. It is widely used for vision tasks in autonomous driving and contains 3000 training images and 500 test images. ADE20K and COCO-stuff are more challenging datasets because they offer a high diversity of indoor and outdoor scenes; and they have a large number of semantic classes. COCO-stuff has 182 classes while ADE20K has 150 classes. These 3 datasets are the standard benchmark in the supervised image synthesis task. In contrast, in all of the previous works on unpaired GANs, the semantic image synthesis experiments were only performed on Cityscapes, due to the challenging nature of COCO-stuff and ADE20K. In what follows, we start by providing more details about the training setup and evaluation metrics. Then, we show the results of our ablation study on Cityscapes to illustrate the role of different parts of the proposed model. Next, we discuss the performance of USIS against the state-of-the-art unpaired models and some of the supervised frameworks. Finally, we showcase the performance of our model in a practical use case: we perform the translation between labels extracted from a modern computer game (GTA-V) and images captured in real-time (Cityscapes).

6.1. Training Details and Evaluation Metrics

We follow BigGAN[26] and OASIS[2] and perform our experiments using an exponential moving average of the generator weights with 0.9999 decay. The image resolution is 256×256 for COCO and ADE20K, and 256×512 for Cityscapes. We use a batchsize of 8 on one Titan-RTX GPU for Cityscapes and a batchsize of 32 for ADE20K and COCO on 4 Titan-RTX GPUs. The optimizer in all our experiments is ADAM[79] with momentums $\beta_1 = 0, \beta_2 = 0.999$ and a learning rate of 0.0001. In Eq. 2, the segmentation loss coefficient λ_{seg} is set to 1.0.

Metrics. A good image generation should satisfy a high diversity of images, a high quality and a multimodal generation (defined as the ability to generate different images from the same label). To this end, the standard evaluation metrics for this task are utilized to measure both the quality and diversity of generated images. The metrics we use are:

- **FID**: We show the Frechet Inception Distance or FID [80],

Method	Implementation Details							256 × 512	
	A ⟶ B	B ⟶ A	Discriminator			Generator		FID	mIoU
			Patch	Whole	Wavelet	CNN	OASIS		
A- CycleGAN[8]	$\checkmark_{L_1}$	✓	✓			✓		87.2	24.5
B- OASIS-CycleGAN	$\checkmark_{L_1}$	✓	✓				✓	147.7	16.9
C- USIS (Patch)	✓		✓				✓	128.67	31.41
D- USIS (Whole)	✓			✓			✓	55.57	35.17
E- USIS (Wavelet)	✓				✓		✓	**52.19**	42.8
F- USIS	✓				$\checkmark_{decoder}$		✓	53.67	**44.78**

Table 1: Ablation study conducted to highlight the impact of the proposed self-supervised training paradigm as well as the selected discriminator and generator architectures. $A \longrightarrow$ B refers to the first cycle loss, from label to image, while $B \longrightarrow$ A refers to the reverse cycle, from images to labels.

to assess both quality and diversity.

- **mIoU**: We also follow the SPADE evaluation protocol [4] and we run pretrained semantic segmentation models on the generated images and report the mean Intersection-over-Union (mIoU) to evaluate the semantic alignment. We employ DRN-D-105[81] (pretrained on multiple scales) for Cityscapes, DeepLabV2[82] for COCO-stuff and UperNet101[83] for ADE20K.

- **MS-SSIM**: to measure the variation of multimodal images produced by the generator, we measure the Multi-scale SSIM (MS-SSIM) between images generated from the same label maps, following OASIS [2]. For each label, we generate 20 images and measure pairwise MS-SSIM between them, and take the average. Lower MS-SSIM implies the generator is able to produce more diverse images from the same label.

Note that in contrast to I2I translation frameworks, we cannot use MS-SSIM, PSNR, MAE or RMSE with respect to the groundtruth images. This is due to the nature of multimodal generation: the network should produce many plausible images from the same labels, not just one image similar to the groundtruth. A high PSNR would mean a worse multimodal generation, which is harmful to the nature of the application.

The reported mIoU is not only a measure of the semantic alignment but also a measure of the quality of the generated images, because even if the image is aligned with the mask but some objects have an unrealistic or an out-of-distribution texture, a pretrained segmentation network will attribute the wrong class to the object. This is mainly due to the bias of the segmentation networks towards the texture or pixel statistics of the input image. Finally, we show qualitative results to showcase the ability of the model to generate multimodal images.

6.2. Ablation Study

Main Ablation. In Table 1, we perform an ablation study on Cityscapes to analyze the effect of the different components on the generation capability of the model. For a fair comparison, all models were trained with the same OASIS generator (SPADE generator with added 3D noise tensor) and a batchsize of 8. We perform our experiments on a resolution of 256×512.

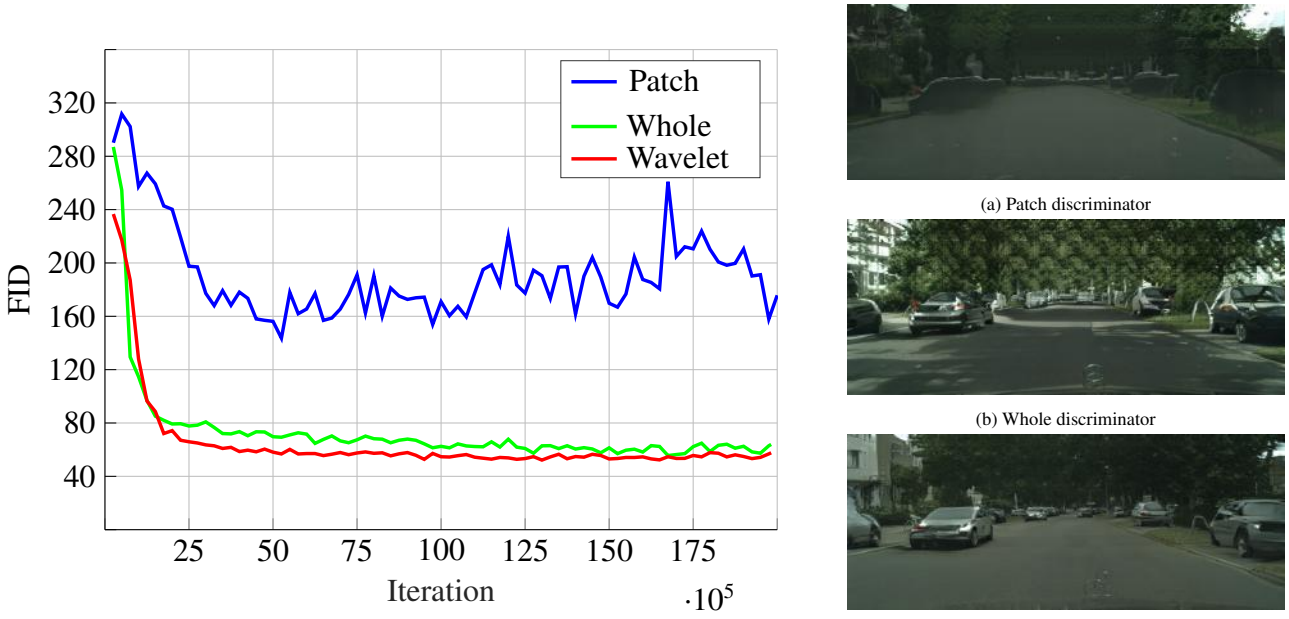
We start with CycleGAN in the first row, which shows an FID of 87.2 and a mIoU of 24.5. Then, in configuration B, we replace the CycleGAN generator with the OASIS generator, and one-hot encode all labels. However, we observe an immediate deterioration of both metrics, showing that a straightforward extension of CycleGAN is not sufficient. In configuration C, we remove the second cycle loss and the second discriminator, replacing the first cycle loss with our self-supervised class-balanced segmentation loss. Immediately, we see a considerable improvement in mIoU, surpassing CycleGAN. The FID decreases but remains high. Hypothesizing that the patch discriminator's small receptive field of view is the main bottleneck, we replace the patch discriminator with a whole image discriminator in configuration D. The effect is most visible in FID, surpassing the state-of-the-art (CUT[1]). Adding the whole discriminator produces images with an overall more realistic texture because the absence of the averaging effect on the output, previously present in the patch discriminator, has enabled starker discrimination.

Note that CUT [1] has also utilized a whole image discriminator in their ablation studies and reported only a marginal increase in the performance. In contrast, the discriminator's receptive field in USIS has a crucial effect on the model's convergence and image quality. Motivated by this finding, we also reason that the discriminator should observe high-frequency bands of the image. In configuration E, we experiment with the wavelet discriminator and notice an improvement in both metrics. The model with the designed decoder in configuration F reaches an FID of 53.67 and a mIoU of 44.78. Overall, this is an improvement of 36.8% in FID and 80.8% in mIoU compared to the CycleGAN baseline.

Ablation study on the discriminator type. In Figure 5, we showcase the isolated effect of the discriminator type on the model training and the image quality. Whole discrimination adds more stability to the training compared to the patch discriminator. On the other hand, wavelet discrimination is slightly slower than whole discrimination but reaches a lower FID and exhibits the same stability during training. In terms of visual quality, the patch discriminator has the worst performance of the three types. Although it is able to match the colors of the bigger classes in Figure 5 (road, tree), the cars are barely visible and have the same color as the road. Although the car bound-

Discriminator	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
Patch	84.9	53.1	59.0	0.2	3.5	36.3	15.2	35.5	56.0	33.5	77.4	44.5	24.4	18.7	0.5	0.0	0.0	15.9	38.0	31.41
Whole	93.2	58.6	65.1	13.1	12.4	31.3	3.3	22.3	65.3	49.1	76.2	35.5	16.1	84.2	10.3	2.2	0.0	1.4	28.6	35.17
Wavelet	92.7	56.3	80.7	20.6	27.8	33.9	21.3	28.9	79.6	57.7	77.6	51.1	34.4	81.1	15.5	11.1	0.01	0.02	43.2	42.80
Wavelet + decoder	93.3	59.9	82.5	27.5	20.0	34.3	23.0	34.8	81.4	58.1	83.8	50.8	23.1	82.6	30.2	21.5	0.00	3.7	40.1	44.78
OASIS[2]	97.1	80.0	85.8	70.0	65.3	40.8	46.1	57.4	85.6	70.2	91.6	63.6	49.9	88.8	78.4	78.4	66.4	47.9	60.7	69.7

Table 2: IoU per class on Cityscapes dataset for the three investigated discriminator architectures in the USIS framework. OASIS is included as an upper bound.



(a) Patch discriminator

(b) Whole discriminator

(c) Wavelet discriminator

Fig. 5: Quantitative and qualitative ablation results of different discriminator architectures on Cityscapes dataset.

aries are discernible (thanks to the self-supervised loss), the texture is almost non-existent. On the other hand, the whole discriminator can generate cars with sharper details (wheels, car lights) but doesn't generate trees with a realistic texture, due to the loss of higher frequencies in the downsampling layers of the generator. The wavelet discriminator solves this issue and combines the advantages of patch and whole image discriminators to yield a higher quality. The effect is visible in both large (street, building and trees) and small classes (cars).

In order to reliably measure the quality of image synthesis, and identify the strengths and weaknesses of the proposed framework, we measure the IoU for each class in the generated Cityscapes images. The IoU was obtained by applying a pre-trained DRN-D-105[81] on Cityscapes, as previously discussed. For the purpose of this study, Cityscapes presents itself as the most suitable dataset because it has a limited amount of classes and the experiments were run on the higher resolution (256×512), making it easier to visualize the effects of the different discriminator architectures. We present our results in Table 2. The improvement brought by the whole discriminator is visible in large and medium classes. Road, sidewalks, buildings and terrain for instance are generated with a better texture. The largest improvement was exhibited by the class car, whose IoU jumped from 18.7 to 84.2. The downside of the whole discriminator is that smaller classes (person, rider, traffic lights and bikes) have suffered from a slight drop in the IoU scores. However, the overall improvement overshadows the negative side effects and lead to an overall better mIoU and FID, which correlates to a better visual perception. The wavelet discriminator keeps the same improvements brought by the whole discriminator and even corrects its shortcomings, by generating the small classes. Compared to the patch discriminator, the wavelet discriminator can generate classes that were almost absent in the images generated by patch discrimination (like wall, truck and bus). The addition of the decoder boosts the IoU of the majority of the classes, especially the rare classes like bus and truck because the discriminator is now forced to reconstruct their wavelet coefficients as well. Finally, we have included OASIS[2] in the comparison as an upper bound to our model, in order to identify improvement opportunities. Most notably, the small classes still need improvement but it's the rare classes that have suffered the most in unsupervised training. Classes like truck, bus, train and motorcycle are not present in a lot of training images so they are assigned a low weight in the discriminator loss. In contrast, the classes that have seen the most improvement are almost present in every scene in Cityscapes, such as Buildings, Cars and even the Person class, which counts as a small class but is frequently seen in the dataset. This pattern is also present in other datasets like ADE20K and Cocostuff, which will be discussed in the Section 6.3. From this series

Config.	Method	FID↓	mIoU↑	MS-SSIM↓
	USIS (no decoder)	52.2	**42.8**	0.51
A	USIS with L1-loss	65.6	31.4	**0.47**
B	USIS with GC [11]	309.6	0.37	1.0
C	USIS + reverse cycle	69.5	41.0	0.54
D	USIS + reverse cycle w/o 3D noise	**49.1**	42.6	1.0

Table 3: Ablation Study on the learning scheme. All configurations have a wavelet discriminator with encoder only. UNet is used in all configurations except in configuration B.

Config.	Method	FID↓	mIoU↑
	USIS (no decoder)	52.2	42.80
A	decoder with Transposed Convs and BN	52.5	41.85
B	Decoder with WU and BN	**50.6**	40.32
C	decoder with WU and IN	53.7	**44.78**

Table 4: Ablation Study on the decoder architecture.

of experiments, we can conclude that the discriminator plays a pivotal role in the learning scheme in unsupervised label-to-image translation, in contrast to unpaired I2I where it only has incremental effects (as mentioned in CUT [1]).

Ablation on the supervision paradigm. After discussing the impact of the discriminator design on image quality, we perform ablation studies on the supervision scheme. Specifically, we seek to answer the question: are other unsupervised learning paradigms [8, 10, 1, 9, 12, 11] constrained to use color-coded semantic labels as input? And how would they perform if we use the same generator and discriminator as USIS? We can answer the first question by examining the architecture of previous works. Notably, some learning paradigms cannot use a one-hot encoded representation for the label. For instance, DRIT [10], and CUT [1] use a shared encoder for input and output domains, which constrain the labels to be color-coded rather than one-hot encoded. MUNIT [9] encodes the input into a content code and a style code, a concept which is undefined for labels and is incompatible with the decoder-like structure of the SPADE generator. DistanceGAN [12] maintains the pixelwise L1-distance between a pair of samples, a concept that works with images but is ill-posed for labels because the L1-loss between images is sensitive to color information, which is absent in one-hot encoded labels. On the other hand, CycleGAN [8] and GCGAN [11] do not exhibit any constraints in their architectures, which enables the use of a one-hot encoded representation. To answer the second question, we extend CycleGAN and GCGAN to use the OASIS generator and the wavelet discriminator. In Table 3, we report the results of various models on the Cityscapes dataset. In Configuration A, we replace the self-supervised segmentation loss with the CycleGAN L1-loss with respect to one-hot encoded labels. In Configuration B, we use the geometric consistency loss (vertical flip) from GCGAN [11] instead of the one-sided cycle loss. A whole discriminator replaces the UNet and is used to discriminate the vertically flipped images. In Configuration C, a reverse cycle is added to USIS similar to CycleGAN (images $\longrightarrow$ labels $\longrightarrow$ images) along with an additional discriminator for labels. Finally, in Configuration D, we remove the 3D noise from the generator with the reverse cycle.

The results show that the L_1-loss leads to worse results (high FID and low mIoU) and that the self-supervised segmentation loss is more suited to preserve semantic consistency. Surprisingly, the Geometric Consistency loss (USIS + GC) fails to converge. We attribute this to the fact that label-to-image translation is more challenging than I2I and that the learning signal from the GC loss is not robust enough when the input is a label map. Adding a reverse cycle to USIS boosts the mIoU to a similar score to USIS. However, the FID is significantly degraded. We attribute this effect to the presence of 3D noise. We hypothesize that the reverse cycle imposes a one-to-one mapping, which is incompatible with 3D noise. In configuration D, we remove the 3D noise from the generator and notice a considerable improvement in FID. However, this comes at the expense of multimodality (MS-SSIM = 1.0, which means no controllable generation is possible). The generator only produces a deterministic image from a label, which is major drawback in this model. This ablation study shows that the unsupervised scheme in USIS can reach an optimal trade-off between quality, diversity, and multimodality compared to other paradigms.

Ablation on the decoder design We study the effect of different layers in the decoder on the performance in Table 4. Adding a traditional decoder leads to a 1 mIoU drop. Replacing transposed convolutions by WU lead to a better FID although the mIoU drops by one further point. Hypothesizing that the batch normalization layers affect the training stability, we replace them by instance normalization layers and directly observe a significant improvement in the mIoU with a negligible drop in the FID.

6.3. Main Results

In Table 5, the performance of USIS is compared against the state-of-the-art models in unpaired I2I translation on all 3 datasets. The results showcase the effectiveness of the proposed framework in the task of unpaired image synthesis, with respect to both FID and mIoU. We incorporate supervised baselines in the table as an upper bound to the unsupervised model and to illustrate that the proposed USIS is a first step to bridge the existing performance gap.

First, it is visible that many unpaired frameworks have better FID scores than paired frameworks, especially the baselines that appeared before SPADE[4]. FID is influenced by both image quality and diversity; and since many supervised baselines were not trained to learn multimodal generation, their FID score is affected. However, they still exhibit a better mIoU, thanks to the supervised losses which establish clear correspondences between input and output. In this case, the higher mIoU score can be interpreted to correlate with a superior visual quality in more classes.

Second, we find that CycleGAN and CUT perform consistently good on the 3 datasets relative to the other baselines. However, we observe in Figure 6 that images were generated with either unrealistic color (in the case of CycleGAN) or unrealistic texture (CUT). Other baselines can more or less approximate the color and texture distributions of real images but often fail to output semantically meaningful objects. For instance, DistanceGAN has a better FID than CycleGAN but it doesn't

Method	Supervised	Cityscapes		ADE20K		COCO-stuff	
		FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CycleGAN[8]	✗	87.2	24.5	96.3	5.4	104.7	2.08
MUNIT[9]	✗	84	8.2	n/a	n/a	n/a	n/a
DRIT[10]	✗	164	9.5	132.2	0.016	135.5	0.008
DistanceGAN[12]	✗	78	17.6	80	0.035	92.4	0.014
GCGAN[11]	✗	80	8.4	92	0.07	99.8	0.019
CUT[1]	✗	57.3	29.8	79.1	6.9	85.6	2.21
USIS	✗	**53.7**	**44.8**	**33.2**	**17.38**	**27.8**	**14.06**
CRN[32]	✓	104.7	52.4	73.3	22.4	70.4	23.7
SIMS[84]	✓	49.7	47.2	n/a	n/a	n/a	n/a
Pix2pixHD[3]	✓	95.0	58.3	81.8	20.3	111.5	14.6
SPADE[4]	✓	71.8	62.3	33.9	38.5	22.6	37.4
CC-FPSE[37]	✓	54.3	65.5	31.7	43.7	19.2	41.6
OASIS[2]	✓	47.7	69.3	28.3	48.8	17.0	44.1

Table 5: Comparison against SOTA methods on 3 datasets. Bold denotes the best performance for unsupervised models while red denotes the state-of-the-art for supervised models and is considered as the upper bound. All models have been evaluated using the officially published codes. GCGAN with the vertical flip consistency was evaluated on Cityscapes because of the rectangular resolution of the images in the dataset while GCGAN with rotation consistency was evaluated on ADE20K and Cocostuff.

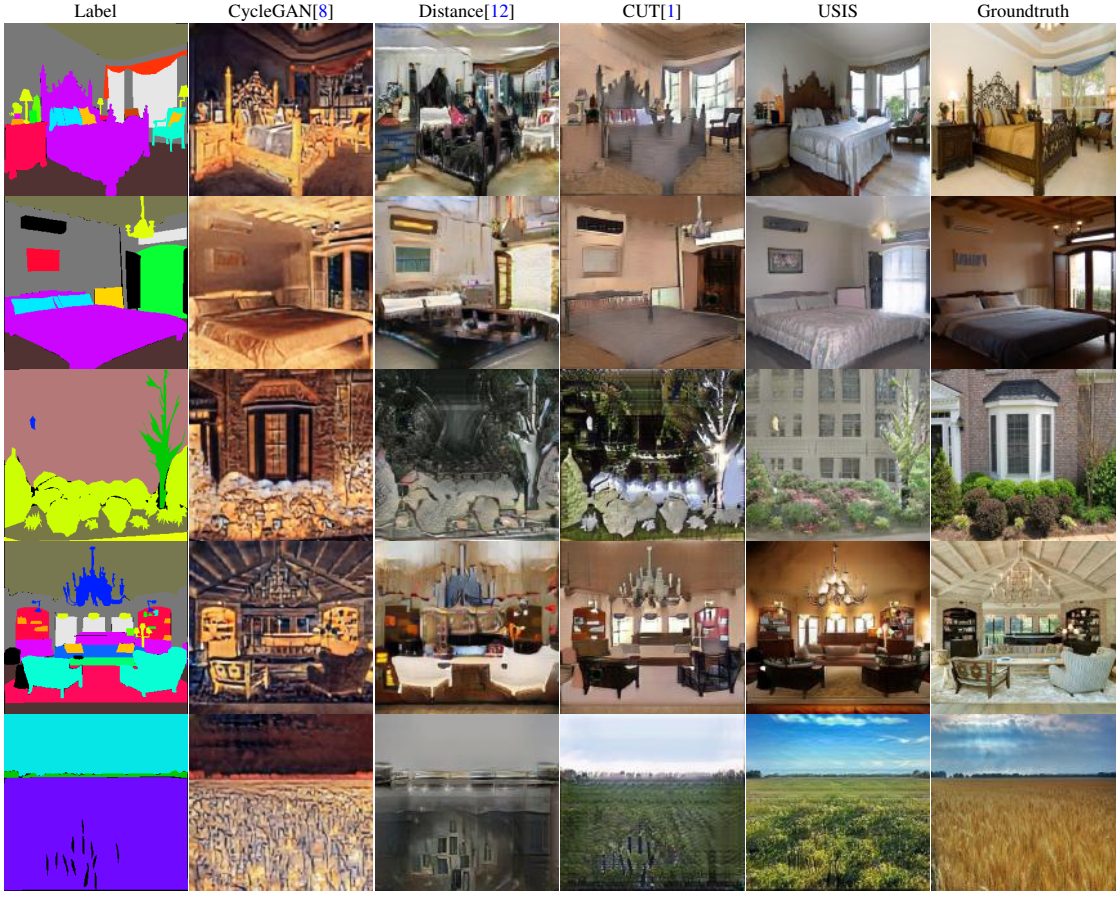


Fig. 6: Qualitative comparison against unpaired GAN frameworks on ADE20K dataset

generate objects with discernible boundaries. In contrast, CycleGAN produces more visible objects (higher mIoU scores) but with unrealsitic color and texture due to the restrictions of the cycle losses. CUT consistently performs better than other baselines.

However, we observe a performance gap between Cityscapes on one hand and ADE20K and Cocostuff on the other hand. To our knowledge, there has been no evaluation of unpaired I2I translation baselines on ADE20K and Cocostuff, so we used the published codes to train and evaluate the aforementioned baselines on these 2 datasets. We assume that the performance drop on ADE20K and Cocostuff correlates directly to the larger number of classes (> 150) they contain in contrast to Cityscapes (34). This entails 2 consequences: first, the color codes assigned to the classes are much closer in value leading most of the baselines to establish label-to-image correspondences that

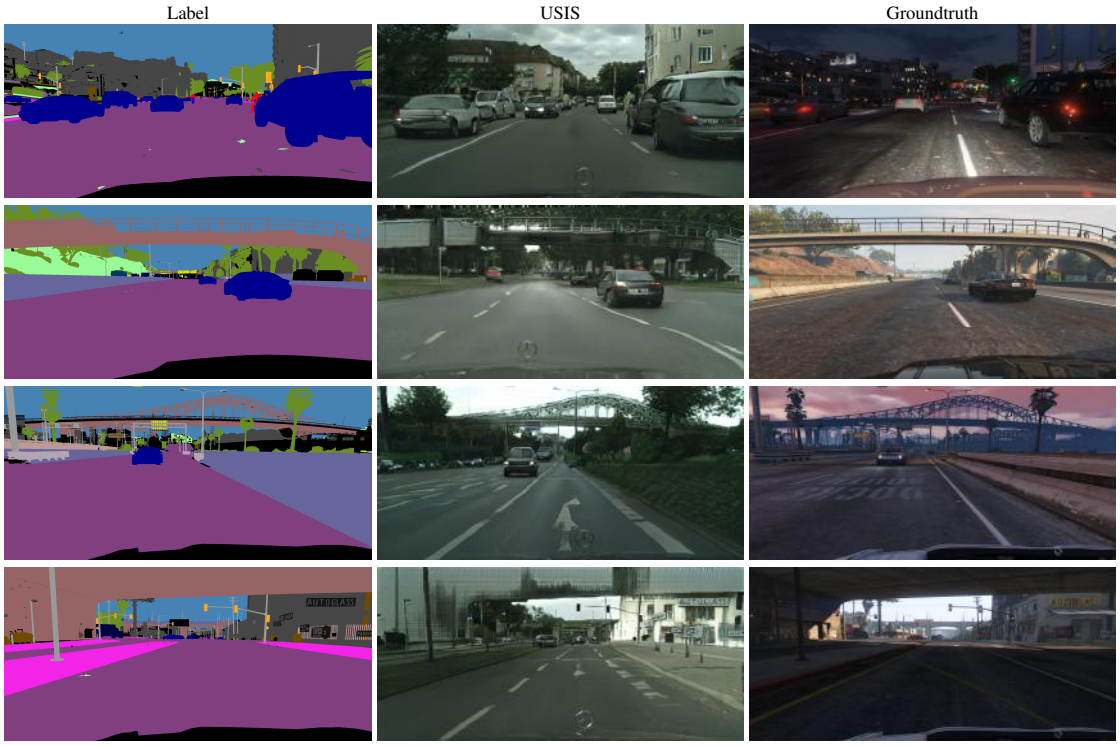


Fig. 7: Results of USIS when trained on GTA-V labels and Cityscapes images

do not preserve semantics; and second, because of the high diversity in images in both datasets (showing indoor and outdoor scenes), there exists a larger number of "rare" classes that should be evaluated by the discriminator. The low mIoU scores in ADE20K and Cocostuff columns quantifies the suboptimal quality observed visually in Figure 6. The proposed model USIS was able to generate more photorealistic classes in many scenes, within indoor and outdoor scenarios. We encourage interested readers to refer to the Appendix for more visual results on the 3 datasets.

6.4. Application on different datasets

Finally, we conduct an experiment for a practical use case where the labels and the images come from two different datasets: GTA-V[13] and Cityscapes. GTA-V is a collection of 25,000 frames taken from a modern computer game and for which dense pixel-level annotations have been generated in an automatic way. In Figure 7, we show some synthesized Cityscapes-like images from GTA-V labels. The task is more challenging because there exists a domain gap between GTA-V labels and Cityscapes labels: on one hand, some elements are present in GTA-V labels but almost absent in Cityscapes (like bridges, tunnels or some steel structures from construction sites); on the other hand, GTA-V is modeled after USA cities while Cityscapes was acquired in German cities leading to a different scene composition. Nevertheless, we have included some challenging scenes in Figure 7 in order to test the performance of the proposed model in such a setting. The results show a texture that is still similar to Cityscapes albeit with more semantic misalignment at the borders of challenging objects. The new objects themselves (bridge, tunnel) are rendered in a reasonable way although they have not been observed by the discriminator during the training. We argue that a simple data augmentation for Cityscapes images (which is easy to acquire) would be enough to enhance the performance. No data augmentation for the labels is needed.

6.5. Limitations and Outlook

Our work is not without limitations. Although USIS outperforms existing unpaired GANs, it is influenced by dataset biases. For instance, if a class is misrepresented in the dataset, USIS might not be able to match its texture. This problem also exists in the supervised setting but is even more challenging when no labels exist. Other dataset biases can have a negative effect on the performance like semantic domain gaps, leading to slightly more misalignment as shown in the previous discussion (section 6.4).

Our framework featured a UNet segmentation network as part of the one-sided cycle loss. However, the effect of this network architecture on the model's learning has not been studied. We believe this has a small effect on the learning, therefore it is out of the scope of this paper and is considered future works.

7. Conclusion

We propose a framework, USIS, for semantic image synthesis in an unpaired setting. It deploys a SPADE generator along with a UNet and an unconditional wavelet-based whole image discriminator with a decoder. The UNet fosters class separability and content preservation while the discriminator matches the color and texture distribution of real images. The decoder prevents the discriminator from overfitting on a few wavelet coefficients. The effectiveness of USIS in the semantic image synthesis was shown on 3 challenging datasets: Cityscapes,

ADE20K and Cocostuff. The proposed framework achieved a significant improvement over prior unsupervised approaches while approaching the performance benchmarks of supervised SIS frameworks. Ablation studies have demonstrated that the discriminator design plays a more important role in this task and can make the difference between non-convergence and a high quality image synthesis. We have also shown that previously designed unsupervised schemes do not extend readily to SIS, and that the supervision paradigm should be adapted to the one-hot encoded representation. Finally, we tested USIS on (GTA-V label)-to-(Cityscapes image) translation to validate its performance in a more challenging setting, as there exists a domain gap between the input labels and the groundtruth images.

USIS is a first step towards bridging the performance gap between paired and unpaired settings. We hope this work would encourage the development of more unsupervised paradigms tailored for label-to-image translation. Furthermore, the research in the unsupervised setting paves the way for the semi-supervised setting which is a promising setup for its practical use cases and the fast improvements it can bring to SIS models while reducing the needed amount of labeled data.

Acknowledgments

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project ”KI Delta Learning – Development of methods and tools for the efficient expansion and transformation of existing AI modules of autonomous vehicles to new domains.” The authors would like to thank the consortium for the successful cooperation.

References

[1] Park, T, Efros, AA, Zhang, R, Zhu, JY. Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision. 2020,.

[2] Schönfeld, E, Sushko, V, Zhang, D, Gall, J, Schiele, B, Khoreva, A. You only need adversarial supervision for semantic image synthesis. In: International Conference on Learning Representations (ICLR). 2021,.

[3] Wang, TC, Liu, MY, Zhu, JY, Tao, A, Kautz, J, Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2018,.

[4] Park, T, Liu, MY, Wang, TC, Zhu, JY. Semantic image synthesis with spatially-adaptive normalization. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2019,.

[5] Pharr, M, Jakob, W, Humphreys, G. Physically Based Rendering: From Theory to Implementation. 3rd ed.; San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2016. ISBN 0128006455.

[6] Sankaranarayanan, S, Balaji, Y, Jain, A, Lim, SN, Chellappa, R. Learning from synthetic data: Addressing domain shift for semantic segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018;:3752–3761.

[7] Cordts, M, Omran, M, Ramos, S, Rehfeld, T, Enzweiler, M, Benenson, R, et al. The cityscapes dataset for semantic urban scene understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2016,.

[8] Zhu, JY, Park, T, Isola, P, Efros, AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International Conference on Computer Vision (ICCV). 2017,.

[9] Huang, X, Liu, MY, Belongie, S, Kautz, J. Multimodal unsupervised image-to-image translation. In: European Conference on Computer Vision (ECCV). 2018,.

[10] Lee, HY, Tseng, HY, Huang, JB, Singh, MK, Yang, MH. Diverse image-to-image translation via disentangled representation. In: European Conference on Computer Vision (ECCV). 2018,.

[11] Fu, H, Gong, M, Wang, C, Batmanghelich, K, Zhang, K, Tao, D. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2019,.

[12] Benaim, S, Wolf, L. One-sided unsupervised domain mapping. In: Advances in Neural Information Processing Systems (NeurIPS). 2017,.

[13] Richter, SR, Vineet, V, Roth, S, Koltun, V. Playing for data: Ground truth from computer games. In: Leibe, B, Matas, J, Sebe, N, Welling, M, editors. European Conference on Computer Vision (ECCV); vol. 9906 of *LNCS*. Springer International Publishing; 2016, p. 102–118.

[14] Geiger, A, Lenz, P, Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2012,.

[15] Gal, R, Cohen, D, Bermano, AH, Cohen-Or, D. Swagan: A style-based wavelet-driven generative model. ArXiv 2021;abs/2102.06108.

[16] Caesar, H, Uijlings, J, Ferrari, V. Coco-stuff: Thing and stuff classes in context. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2018,.

[17] Zhou, B, Zhao, H, Puig, X, Fidler, S, Barriuso, A, Torralba, A. Scene parsing through ade20k dataset. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2017,.

[18] Goodfellow, I, Pouget-Abadie, J, Mirza, M, Xu, B, Warde-Farley, D, Ozair, S, et al. Generative adversarial nets. In: NIPS. 2014,.

[19] Karras, T, Aila, T, Laine, S, Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. ArXiv 2018;abs/1710.10196.

[20] Karras, T, Laine, S, Aila, T. A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019;:4396–4405.

[21] Huang, X, Belongie, SJ. Arbitrary style transfer in real-time with adaptive instance normalization. 2017 IEEE International Conference on Computer Vision (ICCV) 2017;:1510–1519.

[22] Karras, T, Laine, S, Aittala, M, Hellsten, J, Lehtinen, J, Aila, T. Analyzing and improving the image quality of stylegan. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020;:8107–8116.

[23] Liu, B, Zhu, Y, Song, K, Elgammal, A. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In: International Conference on Learning Representations. 2020,.

[24] Mescheder, LM, Geiger, A, Nowozin, S. Which training methods for gans do actually converge? In: ICML. 2018,.

[25] Miyato, T, Koyama, M. cGans with Projection Discriminator. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net; 2018,URL: https://openreview.net/forum?id=ByS1VpgRZ.

[26] Brock, A, Donahue, J, Simonyan, K. Large scale gan training for high fidelity natural image synthesis. 2018. arXiv:1809.11096.

[27] Reed, SE, Akata, Z, Yan, X, Logeswaran, L, Schiele, B, Lee, H. Generative adversarial text to image synthesis. In: ICML. 2016,.

[28] Xu, T, Zhang, P, Huang, Q, Zhang, H, Gan, Z, Huang, X, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, p. 1316–1324.

[29] Hong, S, Yang, D, Choi, J, Lee, H. Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, p. 7986–7994.

[30] Isola, P, Zhu, JY, Zhou, T, Efros, AA. Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2017,.

[31] Tang, H, Xu, D, Yan, Y, Torr, PH, Sebe, N. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2020,.

[32] Chen, Q, Koltun, V. Photographic image synthesis with cascaded refinement networks. In: International Conference on Computer Vision (ICCV). 2017,.

[33] Johnson, J, Alahi, A, Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (ECCV). 2016,.

[34] Gatys, LA, Ecker, AS, Bethge, M. Image style transfer using convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2016,.

[35] Simonyan, K, Zisserman, A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). 2015,.

[36] Tan, Z, Chen, D, Chu, Q, Chai, M, Liao, J, He, M, et al. Rethinking spatially-adaptive normalization. ArXiv 2020;abs/2004.02867.

[37] Liu, X, Yin, G, Shao, J, Wang, X, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS). 2019,.

[38] Zhu, P, Abdal, R, Qin, Y, Wonka, P. Sean: Image synthesis with semantic region-adaptive normalization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020;:5103–5112.

[39] Ronneberger, O, Fischer, P, Brox, T. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. 2015,.

[40] Souly, N, Spampinato, C, Shah, M. Semi supervised semantic segmentation using generative adversarial network. In: International Conference on Computer Vision (ICCV). 2017,.

[41] Schönfeld, E, Schiele, B, Khoreva, A. A u-net based discriminator for generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2020,.

[42] Yi, Z, Zhang, H, Tan, P, Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In: International Conference on Computer Vision (ICCV). 2017,.

[43] Choi, Y, Choi, M, Kim, M, Ha, JW, Kim, S, Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2018,.

[44] Liu, MY, Breuel, T, Kautz, J. Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems (NeurIPS). 2017,.

[45] Almahairi, A, Rajeswar, S, Sordoni, A, Bachman, P, Courville, A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In: International Conference on Machine Learning (ICML). 2018,.

[46] Gokaslan, A, Ramanujan, V, Ritchie, D, In Kim, K, Tompkin, J. Improving shape deformation in unsupervised image-to-image translation. In: European Conference on Computer Vision (ECCV). 2018,.

[47] Liang, X, Zhang, H, Lin, L, Xing, E. Generative semantic manipulation with mask-contrasting gan. In: European Conference on Computer Vision (ECCV). 2018,.

[48] Tang, H, Xu, D, Sebe, N, Yan, Y. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: International Joint Conference on Neural Networks (IJCNN). 2019,.

[49] Wu, W, Cao, K, Li, C, Qian, C, Loy, CC. Transgaga: Geometry-aware unsupervised image-to-image translation. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2019,.

[50] Zhang, R, Pfister, T, Li, J. Harmonic unpaired image-to-image translation. In: International Conference on Learning Representations (ICLR). 2019,.

[51] Taigman, Y, Polyak, A, Wolf, L. Unsupervised cross-domain image generation. In: International Conference on Learning Representations (ICLR). 2017,.

[52] Shrivastava, A, Pfister, T, Tuzel, O, Susskind, J, Wang, W, Webb, R. Learning from simulated and unsupervised images through adversarial training. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2017,.

[53] Bousmalis, K, Silberman, N, Dohan, D, Erhan, D, Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2017,.

[54] Amodio, M, Krishnaswamy, S. Travelgan: Image-to-image translation by transformation vector learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019, p. 8983–8992.

[55] Chen, Y, Li, G, Jin, C, Liu, S, Li, TH. Ssd-gan: Measuring the realness in the spatial and spectral domains. In: AAAI. 2021,.

[56] Durall, R, Keuper, M, Keuper, J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020;:7887–7896.

[57] Dzanic, T, Witherden, F. Fourier spectrum discrepancies in deep network generated images. ArXiv 2020;abs/1911.06465.

[58] Bracewell, RN, Bracewell, RN. The Fourier transform and its applications; vol. 31999. McGraw-Hill New York; 1986.

[59] Daubechies, I. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans Inf Theory 1990;36:961–1005.

[60] Daubechies, I. Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics; 1992. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611970104. doi:10.1137/1.9781611970104. arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611970104.

[61] Gao, X, Xiong, H. A hybrid wavelet convolution network with sparse-coding for image super-resolution. 2016 IEEE International Conference on Image Processing (ICIP) 2016;:1439–1443.

[62] Liu, P, Zhang, H, Lian, W, Zuo, W. Multi-level wavelet convolutional neural networks. IEEE Access 2019;7:74973–74985.

[63] Williams, T, Li, RY. Wavelet pooling for convolutional neural networks. In: ICLR. 2018,.

[64] Yoo, J, Uh, Y, Chun, S, Kang, B, Ha, JW. Photorealistic style transfer via wavelet transforms. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) 2019;:9035–9044.

[65] Liu, L, Liu, J, Yuan, S, Slabaugh, G, Leonardis, A, gang Zhou, W, et al. Wavelet-based dual-branch network for image demoireing. ArXiv 2020;abs/2007.07173.

[66] Kang, E, Min, J, Ye, JC. A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. Medical Physics 2017;44:e360–e375.

[67] Liu, Y, Li, Q, Sun, Z. Attribute-aware face aging with wavelet-based generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019;:11869–11878.

[68] Zhang, Q, Wang, H, Du, T, Yang, S, Wang, Y, Xing, Z, et al. Super-resolution reconstruction algorithms based on fusion of deep learning mechanism and wavelet. In: AIPR '19. 2019,.

[69] Wang, J, Deng, X, Xu, M, Chen, C, Song, Y. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video. ArXiv 2020;abs/2008.00499.

[70] Huang, H, He, R, Sun, Z, Tan, T. Wavelet domain generative adversarial network for multi-scale face hallucination. International Journal of Computer Vision 2019;127:763–784.

[71] He, K, Zhang, X, Ren, S, Sun, J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016;:770–778.

[72] Dosovitskiy, A, Brox, T. Generating images with perceptual similarity metrics based on deep networks. In: Lee, DD, Sugiyama, M, Luxburg, UV, Guyon, I, Garnett, R, editors. Advances in Neural Information Processing Systems (NeurIPs). 2016,.

[73] Deng, J, Dong, W, Socher, R, Li, LJ, Li, K, Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009, p. 248–255.

[74] Geirhos, R, Rubisch, P, Michaelis, C, Bethge, M, Wichmann, F, Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. ArXiv 2019;abs/1811.12231.

[75] Zhu, JY, Zhang, R, Pathak, D, Darrell, T, Efros, AA, Wang, O, et al. Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems (NeurIPS). 2017,.

[76] Drucker, H, LeCun, Y. Improving generalization performance using double backpropagation. IEEE transactions on neural networks 1992;3 6:991–7.

[77] Ross, AS, Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: AAAI. 2018,.

[78] Collins, E, Bala, R, Price, B, Süsstrunk, S. Editing in style: Uncovering the local semantics of gans. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020;:5770–5779.

[79] Kingma, DP, Ba, J. Adam: A method for stochastic optimization. In: Bengio, Y, LeCun, Y, editors. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015,URL: http://arxiv.org/abs/1412.6980.

[80] Heusel, M, Ramsauer, H, Unterthiner, T, Nessler, B, Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). 2017,.

[81] Yu, F, Koltun, V, Funkhouser, T. Dilated residual networks. In: Confer-

ence on Computer Vision and Pattern Recognition (CVPR). 2017,.

[82] Chen, LC, Papandreou, G, Kokkinos, I, Murphy, K, Yuille, AL. Semantic image segmentation with deep convolutional nets and fully connected crfs. International Conference on Learning Representations (ICLR) 2015;.

[83] Xiao, T, Liu, Y, Zhou, B, Jiang, Y, Sun, J. Unified perceptual parsing for scene understanding. In: European Conference on Computer Vision (ECCV). 2018,.

[84] Qi, X, Chen, Q, Jia, J, Koltun, V. Semi-parametric image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2018,.

Supplementary Material

In Figure 8, we showcase the ability of our model to generate multimodal images by sampling several times from the 3D noise at the input of the generator. We perform this experiment on ADE20K. In Figures 9, 10 and 11, we show more qualitative results of our model against other baselines, on Cityscapes, Cocostuff and ADE20K, respectively.

Fig. 8: Multimodal generation on ADE20K by resampling from the 3D noise

Fig. 9: Qualitative comparison on Cityscapes dataset

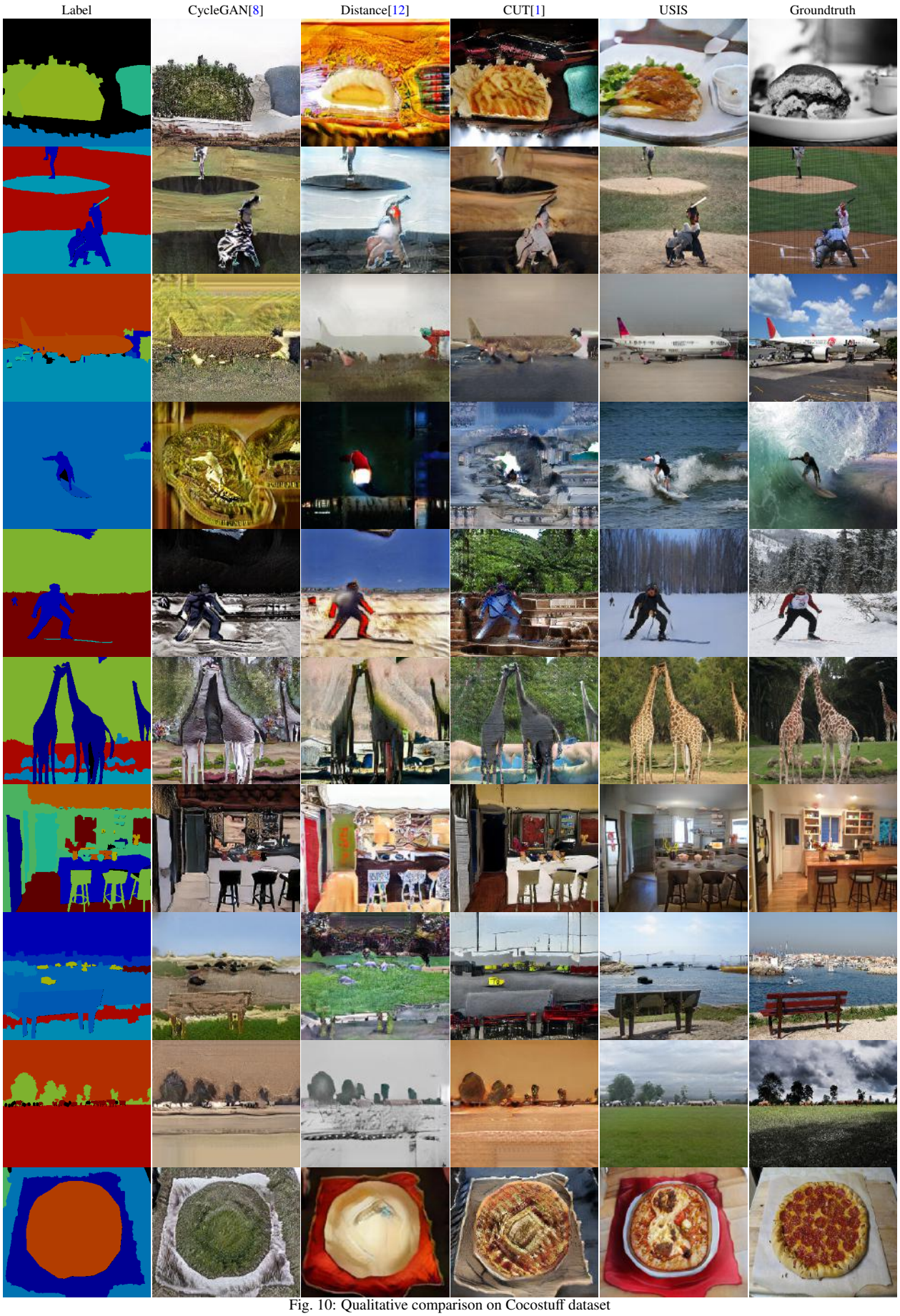


Fig. 10: Qualitative comparison on Cocostuff dataset

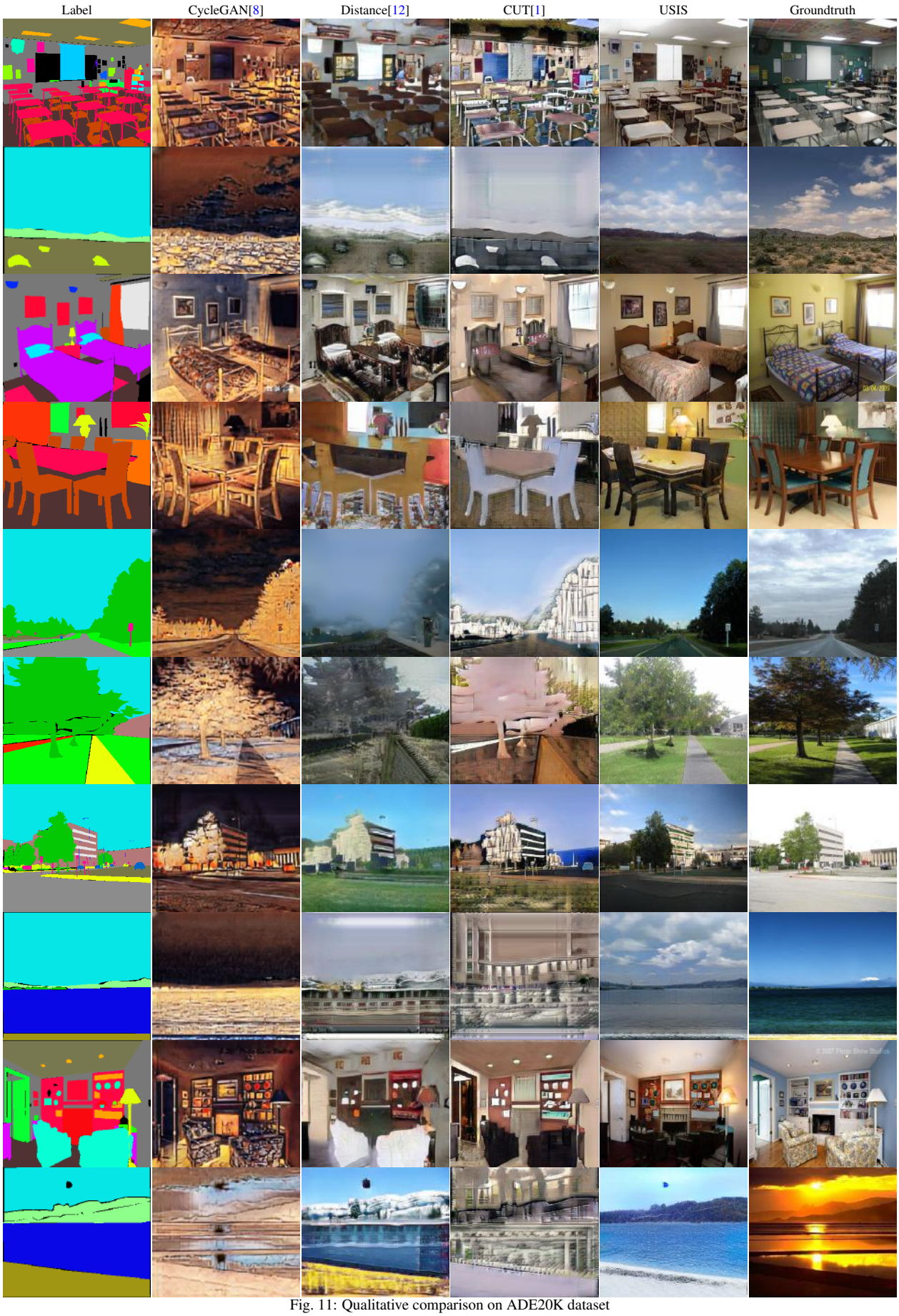


Fig. 11: Qualitative comparison on ADE20K dataset