Semantic-aware Grad-GAN for Virtual-to-Real Urban Scene Adaption

Peilun Li[1]
peilunl@andrew.cmu.edu
Xiaodan Liang[1]
xiaodan1@cs.cmu.edu
Daoyuan Jia[1]
daoyuanj@andrew.cmu.edu
Eric P. Xing[2]
epxing@cs.cmu.edu

[1] School of Computer Science
Carnegie Mellon University
Pittsburgh, USA

[2] Petuum, Inc. Pittsburgh, USA

Abstract

Recent advances in vision tasks (e.g., segmentation) highly depend on the availability of large-scale real-world image annotations obtained by cumbersome human labors. In this work, we resort to transfer knowledge from automatically rendered scene annotations in virtual-world to facilitate real-world visual tasks. Although virtual-world annotations can be ideally diverse and unlimited, the discrepant data distributions between virtual and real-world make it challenging for knowledge transferring. We thus propose a novel Semantic-aware Grad-GAN (SG-GAN) to perform virtual-to-real domain adaption with the ability of retaining vital semantic information. Beyond the simple holistic color/texture transformation achieved by prior works, SG-GAN successfully personalizes the appearance adaption for each semantic region in order to preserve their key characteristic for better recognition. Qualitative and quantitative experiments demonstrate the superiority of SG-GAN in scene adaption over state-of-the-art GANs. Further evaluations on semantic segmentation on Cityscapes show using adapted virtual images by SG-GAN dramatically improves segmentation performance than original virtual data. We release our code at https://github.com/Peilun-Li/SG-GAN.

1 Introduction

Recently, very promising visual perception performances on a variety of tasks (e.g., classification and detection) have been achieved by deep learning models, driven by large-scale annotated datasets. However, more fine-grained tasks (e.g., semantic segmentation) still have much space to be resolved due to the insufficient pixel-wise annotations. High quality annotations are often prohibitively difficult to obtain with the need of tons of human efforts. An alternative solution to alleviate this data issue is to seek an automatic data generation approach. Rather than relying on expensive labors on annotating real-world data, recent progresses [16, 24, 25] make it possible to automatically capture both images and semantic labeling in an unlimited way from video games, e.g., GTA V. However, utilizing virtual-world knowledge to facilitate real-world perception tasks is not a trivial technique since images

(a) Real-world (Cityscapes)

(b) Virtual-world (GTA-V)

Figure 1: Visual comparison between real-world images and virtual-world images. (a) Real-world images (Cityscapes dataset) (b) Virtual-world images (GTA-V dataset)

collected from virtual-world and real-world are sampled from different underlying distributions, as shown in Figure 1. It is thus desirable to bridge the gap between virtual-world and real-world data for exploiting the shared semantic knowledge for perception.

In this work, we propose a novel Semantic-aware Grad-GAN (SG-GAN) that aims at transferring personalized styles (e.g., color, texture) for distinct semantic regions in virtual-world images to approximate real-world distributions. SG-GAN is able to not only preserve key semantic and structure information in source domain but also enforce each semantic region close to their corresponding real-world distributions.

Except the traditional adversarial objective used in prior GANs, we propose two main contributions to achieve the above mentioned goals. First, a new gradient-sensitive objective is introduced into optimizing the generator, which emphasizes the semantic boundary consistencies between virtual images and adapted images. It is able to regularize the generator render distinct color/texture for each semantic region in order to keep semantic boundaries. Second, previous works often learn a whole image discriminator for validating the fidelity of all regions, which makes the color/texture of all pixels in original images easily collapse into a monotonous pattern. We here argue that the appearance distributions for each semantic region should be regarded differently and purposely. In contrast to standard discriminator that eventually examines on a global feature map, we employ a new semantic-aware discriminator for evaluating the image adaption quality in a semantic-wise manner. The semantic-aware discriminator learns distinct discriminate parameters for examining regions with respect to each semantic label. This distinguishes SG-GAN with existing GANs as a controllable architecture that personalizes texture rendering for different semantic regions and results in adapted images with finer details.

Extensive qualitative and quantitative experiments on adapting GTA-V virtual images demonstrate SG-GAN can successfully generate realistic images without changing semantic information. To further demonstrate the quality of adapted images, we use the adapted images to train semantic segmentation models and evaluate them on public Cityscapes dataset [3]. The substantial performance improvement over using original virtual data on semantic segmentation speaks well the superiority of our SG-GAN, and reveals the possibility to apply gradient-sensitive objective and semantic-aware discriminator to other segmentation related models for further boosting their performance.

2 Related work

Real-world vs. virtual-world data acquiring: Fine-grained semantic segmentation on urban scenes takes huge amount of human effort, which results in much less data than that of

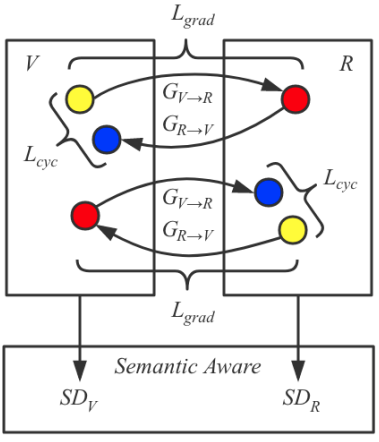


Figure 2: Illustration of SG-GAN.

image classification datasets. For example, Cityscapes dataset [3] releases 5000 road scene annotations and reports annotation speed as more than 90 minutes/image. On the contrary, collecting urban scene data from video games such as GTA V has attracted lots of interests [16, 24, 25] for automatically obtaining a large amount of data. Specifically, Richter *et al*. [24] collect 24966 images with annotations from GTA V within 49 hours. Richter *et al*. [25] further develop real-time rendering pipelines and release a dataset of 254064 fully annotated video frames. However, despite its diversity, virtual-world scene data often looks very unrealistic (e.g., flawed lighting and shadowing) due to the imperfect texture rendering. Directly utilizing such unrealistic data would damage real-world visual tasks due to their discrepant data distributions.

Domain adaption: Domain adaption can be approached by either adapting scene images or adapting hidden feature representations guided by the targets. Image-based adaption can be also referred to as image-to-image translation, i.e., translating images from source domain to target domain, which can be summarized into two following directions.

First, adapted images can be generated through feature matching [6, 7, 15]. Gatys *et al*. [6] propose a method to combine content of one image and style of another image through matching Gram matrix on deep feature maps, at the expense of some loss of content information. Second, a generative model can be trained through adversarial learning for image translation. Isola *et al*. [14] use conditional GANs to learn mapping function from source domain to target domain, with a requirement of paired training data, which is unpractical for some tasks. To remove the requirement of paired training data, extra regularization could be applied, including self-regularization term [26] , cycle structure [17, 32, 33] or weight sharing [21, 22]. There are also approaches making use of both feature matching and adversarial learning [31]. However, in urban scene adaption, despite having the ability to generate relatively realistic images, existing approaches often modify semantic information, e.g., the sky will be adapted to tree structure, or a road lamp may be rendered from nothing.

In contrast to image-based adaption that translates images to target domain, hidden feature representation based adaption adapts learned models to target domain [8, 11, 12, 13, 18, 19, 20, 28]. By sharing weight [8] or incorporating adversarial discriminative setting [28], those feature-based adaption methods help mitigate performance degradation caused by domain shifting. However, feature-based adaption methods require different objective or architecture for different vision tasks, thus not as widely-applicable as image-based adaption.

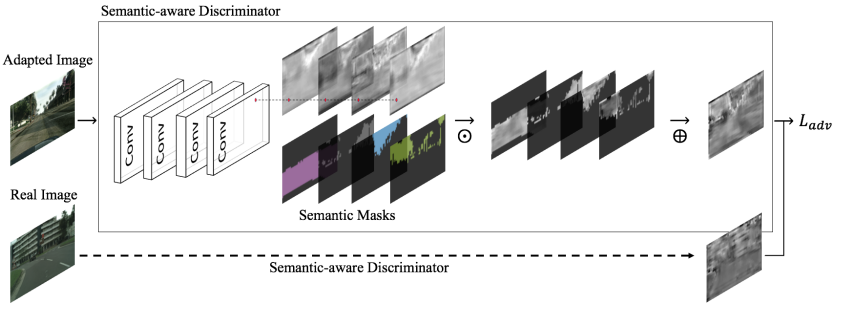


Figure 3: Illustration of semantic-aware discriminator.

3 Semantic-aware Grad-GAN

The goal of the proposed SG-GAN is to perform virtual-to-real domain adaption while preserving their key semantic characteristics for distinct contents. Capitalized on the Generative Adversarial Networks (GANs), SG-GAN presents two improvements over the traditional GAN model, i.e., a new soft gradient-sensitive objective over generators and a novel semantic-aware discriminator.

3.1 Semantic-aware cycle objective

Our SG-GAN is based on the cycle-structured GAN objective since it has shown the advantages of training stability and generation quality [17, 32, 33]. Specifically, let us denote the unpaired images from the virtual-world domain V and real-world domain R as $\{v\}_{i=1}^{N} \in V$ and $\{r\}_{j=1}^{M} \in R$, respectively. Our SG-GAN learns two symmetric mappings $G_{V\to R}$, $G_{R\to V}$ along with two corresponding semantic-aware discriminators SD_R, SD_V in an adversarial way. $G_{V\to R}$ and $G_{R\to V}$ map images between virtual-world and real-world domains. SD_R's target is to distinguish between real-world images $\{r\}$ and fake real-world images $\{G_{V\to R}(v)\}$, and vice versa for SD_V. The details of semantic-aware discriminators will be introduced later in Section 3.2. Figure 2 illustrates the relationship of V, R, $G_{V\to R}$, $G_{R\to V}$, SD_V and SD_R.

3.1.1 Adversarial loss

Our objective function is constructed based on standard adversarial loss [9]. Two sets of adversarial losses are applied to $(G_{V\to R}, SD_R)$ and $(G_{R\to V}, SD_V)$ pairs. Specifically, the adversarial loss L_{adv} for optimizing $(G_{V\to R}, SD_R)$ is defined as:

$$\begin{aligned} L_{adv}(G_{V\to R}, SD_R, V, R) = {} & \mathbb{E}_{r\sim p_{data}(r)}[\log SD_R(r)] \\ & + \mathbb{E}_{v\sim p_{data}(v)}[\log(1 - SD_R(G_{V\to R}(v))] \end{aligned} \tag{1}$$

The formula is similar for the generator $G_{R\to V}$ and semantic-aware discriminator SD_V, of which the adversarial loss can be noted as $L_{adv}(G_{R\to V}, SD_V, R, V)$.

3.1.2 Cycle consistency loss

Another part of our objective function is cycle consistency loss [33], which is shown helpful to reduce the space of possible mappings, i.e., $G_{V\to R}$ and $G_{R\to V}$. In this work, we define

cycle consistency loss as:

$$L_{cyc}(G_{V\to R}, G_{R\to V}, V, R) = \mathbb{E}_{r\sim p_{data}(r)}[||G_{V\to R}(G_{R\to V}(r)) - r||_1] + \mathbb{E}_{v\sim p_{data}(v)}[||G_{R\to V}(G_{V\to R}(v)) - v||_1] \tag{2}$$

For complex adaption such as urban scene adaption, a model purely with cycle consistency loss often fails by wrongly mapping a region with one semantic label to another label, e.g., the sky region may be wrongly adapted into a tree region, as shown in Figure 4. This limitation of cycle structure is also discussed in [33].

3.1.3 Soft gradient-sensitive objective

In order to keep semantic information from being changed through the mapping functions, we introduce a novel soft gradient-sensitive loss, which uses image's semantic information in a gradient level. We first introduce gradient-sensitive loss, and then show ways to make the gradient-sensitive loss into a soft version.

The motivation of gradient-sensitive loss is that no matter how texture of each semantic class changes, there should be some distinguishable visual differences at the boundaries of semantic classes. Visual differences for adjacent pixels can be captured through convolving gradient filters upon the image. A typical choice of gradient filter is Sobel filter [27].

Since our focus is visual differences on semantic boundaries, a 0-1 mask is necessary that only has non-zero values on semantic boundaries. Such mask can be retrieved by convolving a gradient filter upon semantic labeling since it only has different adjacent values on semantic boundaries. By multiplying the convolved semantic labeling and the convolved image element-wise, attention will only be paid to visual differences on semantic boundaries.

More specifically, for an input image v and its corresponding semantic labeling s_v, since we desire v and $G_{V\to R}(v)$ share the same semantic information, the gradient-sensitive loss for image v can be defined as Equation 3, in which $\mathbf{C}_i$ and $\mathbf{C}_s$ are gradient filters for image and semantic labeling, $*$ stands for convolution, $\odot$ stands for element-wise multiplication, $|\cdot|$ represents absolute value, $||\cdot||_1$ means L1-norm, and *nonzero* is a function yielding 1 for nonzero values and 0 otherwise.

$$l_{grad}(v, s_v, G_{V\to R}) = ||(|(|\mathbf{C}_i * v| - |\mathbf{C}_i * G_{V\to R}(v)|)|) \odot nonzero(\mathbf{C}_s * s_v)||_1 \tag{3}$$

In practice, we may hold belief that v and $G_{V\to R}(v)$ share similar texture within semantic classes. Since texture information can also be extracted from image gradient, a soft gradient-sensitive loss for image v can be defined as Equation 4 to represent such belief, in which β controls how much belief we have on texture similarities.

$$l_{s-grad}(v, s_v, G_{V\to R}, \alpha, \beta) = ||(|(|\mathbf{C}_i * v| - |\mathbf{C}_i * G_{V\to R}(v)|)|) \odot (\alpha \times nonzero(\mathbf{C}_s * s_v) + \beta)||_1 \quad s.t.\ \ \alpha + \beta = 1\ \ \alpha, \beta \geq 0 \tag{4}$$

Given the soft gradient-sensitive loss for a single image, the final objective for soft gradient-sensitive loss can be defined as Equation 5, in which S_V is semantic labeling for V and S_R is semantic labeling for R.

$$L_{grad}(G_{V\to R}, G_{R\to V}, V, R, S_V, S_R, \alpha, \beta) = \mathbb{E}_{r\sim p_{data}(r)}[l_{s-grad}(r, s_r, G_{R\to V}, \alpha, \beta)] + \mathbb{E}_{v\sim p_{data}(v)}[l_{s-grad}(v, s_v, G_{V\to R}, \alpha, \beta)] \tag{5}$$

Figure 4: Visual comparison with state-of-the-art methods and our variants. More examples in supplementary materials.

3.1.4 Full objective function

Our full objective function is a combination of adversarial loss, cycle consistency loss and soft gradient-sensitive loss, as Equation 6, where λ_c and λ_g control the relative importance of cycle consistency loss and soft gradient-sensitive loss, compared with adversarial loss.

$$\begin{aligned} &L(G_{V\to R}, G_{R\to V}, SD_V, SD_R) \\ =&L_{adv}(G_{V\to R}, SD_R, V, R) + L_{adv}(G_{R\to V}, SD_V, R, V) \\ +&\lambda_c L_{cyc}(G_{V\to R}, G_{R\to V}, V, R) + \lambda_g L_{grad}(G_{V\to R}, G_{R\to V}, V, R, S_V, S_R, \alpha, \beta) \end{aligned} \tag{6}$$

Our optimization target can be then represented as:

$$G^*_{V\to R}, G^*_{R\to V} = \arg\min_{\substack{G_{V\to R}\ SD_R \\ G_{R\to V}\ SD_V}} \max L(G_{V\to R}, G_{R\to V}, SD_V, SD_R) \tag{7}$$

3.2 Semantic-aware discriminator

The introduction of soft gradient-sensitive loss contributes to smoother textures and clearer semantic boundaries (Figure 5). However, scene adaption also needs to retain more high-level semantic consistencies for each specific semantic region. A typical example is after the virtual-to-real adaption, the tone goes dark for the whole image as real-world images are not as luminous as virtual-world images, however, we may only want roads to be darker without changing much of the sky, or even make sky lighter. The reason for yielding such inappropriate holistic scene adaption is that the traditional discriminator only judges realism image-wise, regardless of texture differences in a semantic-aware manner. To make discriminator semantic-aware, we introduce semantic-aware discriminators SD_V and SD_R. The idea is to create a separate channel for each different semantic class in the discriminator. In practice, this can be achieved by transiting the number of filters in the last layer of standard discriminator to number of semantic classes, and then applying semantic masks upon filters to let each of them focus on different semantic classes.

More specifically, the last (k-th) layer's feature map of a standard discriminator is typically a tensor $\mathbf{T}_k$ with shape $(w_k, h_k, 1)$, where w_k stands for width and h_k stands for height. $\mathbf{T}_k$ will then be compared with an all-one or all-zero tensor to calculate adversarial objective. In contrast, the semantic-aware discriminator we propose will change $\mathbf{T}_k$ as a tensor with shape (w_k, h_k, s), where s is the number of semantic classes. We then convert image's semantic labeling to one-hot style and resize to (w_k, h_k), which will result in a mask $\mathbf{M}$ with same shape (w_k, h_k, s), and $\{\mathbf{M}_{ij}\} \in \{0, 1\}$. By multiplying $\mathbf{T}_k$ and $\mathbf{M}$ element-wise, each

Method A \ Method B	CycleGAN	DualGAN	SimGAN	BiGAN	SG-GAN-2K
SG-GAN-2K	79.2%	93.4%	97.2%	99.8%	—
SG-GAN-25K	83.4%	94.0%	98.4%	99.8%	53.8%

Table 1: Results of A/B tests on Amazon Mechanical Turk (AMT). Each cell shows the proportion that image adapted by method A is chosen as more realistic.

Adapted with L_{grad} Adapted without L_{grad}

Figure 5: 4X zoomed adapted images for showing the effectiveness of L_{grad} objective.

filter within $\mathbf{T}_k$ will only focus on one particular semantic class. Finally, by summing up $\mathbf{T}_k$ along the last dimension, a tensor with shape $(w_k, h_k, 1)$ will be acquired and adversarial objective can be calculated the same way as the standard discriminator. Figure 3 gives an illustration of proposed semantic-aware discriminator.

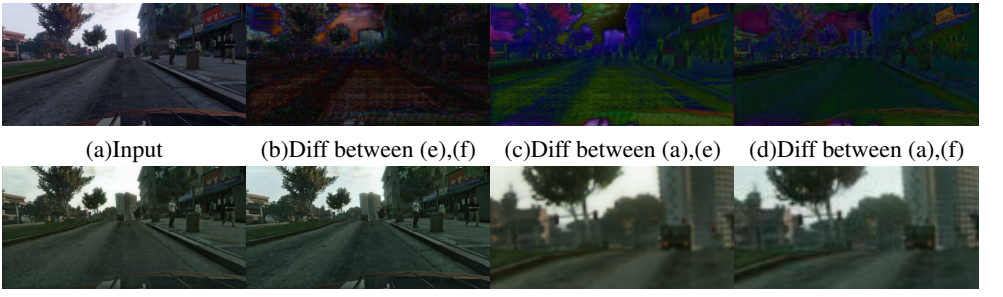
4 Experiments

4.1 Implementation

Dataset. We randomly sample 2000 images each from GTA-V dataset [24] and Cityscapes training set [3] as training images for V and R. Another 500 images each from GTA-V dataset and Cityscapes training set are sampled for visual comparison and validation. Cityscapes validation set is not used for validating adaption approaches here since it will later be applied to evaluate semantic segmentation scores in Section 4.4. We train SG-GAN on such dataset and term it as **SG-GAN-2K**. The same dataset is used for training all baselines in Section 4.2, making them comparable with SG-GAN-2K. To study the effect of virtual-world images, we further expand virtual-world training images to all 24966 images of GTA-V dataset, making a dataset with 24966 virtual images and 2000 real images. A variant of SG-GAN is trained on the expanded dataset and termed as **SG-GAN-25K**.

Network architecture. We use 256×512 images for training phase due to GPU memory limitation. For the generator, we adapt the architecture from [14], which is a U-Net structure with skip connections between low level and high level layers. For the semantic-aware discriminator, we use a variant of PatchGAN [14, 33], which is a fully convolutional network consists of multiple layers of (leaky-ReLU, instance norm [29], convolution) and helps the discriminator identify realism patch-wise.

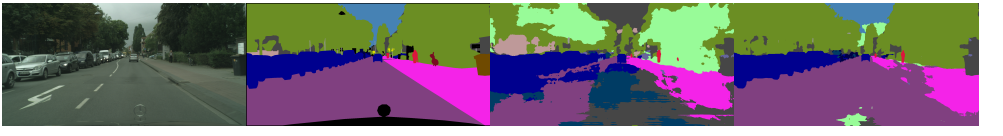
Training details. To stabilize training, we use history of refined images [26] for training semantic-aware discriminators SD_V and SD_R. Moreover, we apply least square objective instead of log likelihood objective for adversarial loss, which is shown helpful in stabilizing training and generating higher quality images, as proposed by Mao *et al*. [23]. For parameters in Equation 6, we set $\lambda_c = 10$, $\lambda_g = 5$. (α, β) is set as $(1, 0)$ for the first three epochs and



(a)Input (b)Diff between (e),(f) (c)Diff between (a),(e) (d)Diff between (a),(f)

(e) Adapted with *SD* (f) Adapted without *SD* (g)4X zoomed (e) (h)4X zoomed (f)

Figure 6: Comparison for showing the effectiveness of semantic-aware discriminator *SD*.



(a) Real-world image (b) Groundtruth (c) Baseline (d) SG-GAN-25K

Figure 7: Comparison of segmentation results, with same color scheme as Cityscapes. More examples in supplementary materials.

then changed to $(0.9, 0.1)$. For gradient filters in Equation 4, we use Sobel filter [27] for $\mathbf{C}_i$ and filters in Equation 8 for $\mathbf{C}_s$ to avoid artifacts on image borders caused by reflect padding. For number of semantic classes in semantic-aware discriminator, we cluster 30 classes [3] into 8 categories to avoid sparse classes, i.e., $s = 8$. Learning rate is set as 0.0002 and we use a batch size of 1. We implement SG-GAN based on TensorFlow framework [1], and train it with a single Nvidia GTX 1080.

$$C_x = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, C_y = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \tag{8}$$

Testing. Semantic information will only be needed at training time. At test time SG-GAN only requires images without semantic information. Since the generators and the discriminators we use are fully convolutional, SG-GAN can handle images with high resolution (*e.g.*, 1024×2048) at test time. The testing time is 1.3 second/image with a single Nvidia GTX 1080.

4.2 Comparison with state-of-the-art methods

We compare our SG-GAN with current state-of-the-art baselines for unpaired virtual-to-real scene adaption for demonstrating its superiority.

4.2.1 Baselines

SimGAN [26] introduces a self-regularization for GAN and local adversarial loss to train a refiner for image adaption. In the experiments we use channel-wise mean values as self-regularization term.

CycleGAN [33] learns mapping functions through adversarial loss and cycle consistency loss. It uses ResNet [10] architecture for the generators and PatchGAN [14] for the discriminators.

Method	Pixel acc.	Class acc.	Class IOU
Baseline	54.51	35.95	24.60
Hoffman *et al*. [13]	–	–	27.10
CycleGAN	71.61	42.98	28.15
SG-GAN-2K	72.65	45.87	33.81
SG-GAN-25K	**81.72**	**47.29**	**37.43**

Table 2: Comparison of semantic segmentation scores (%) on Cityscapes validation set.

DualGAN [32] uses U-Net structure for generators that are identical with SG-GAN. It uses the same PatchGAN structure as CycleGAN, but different from CycleGAN it follows the loss format and training procedure proposed in Wasserstein GAN [2].

BiGAN [4, 5] learns the inverse mapping of standard GANs, which can also be used for unpaired scene adaption.

4.2.2 Qualitative and quantitative evaluation

Figure 4 compares between SG-GAN-2K and other state-of-the-art methods visually. In general, SG-GAN generates better visualization results, in the form of clear boundaries, consistent semantic classes, smooth texture, etc. Moreover, SG-GAN-2K shows its ability for personalized adaption, e.g., while we retain the red color of vehicle's headlight, the red color of sunset is changed to sunny yellow that is closer to real-world images.

To further evaluate our approach quantitatively, we conduct A/B tests on Amazon Mechanical Turk (AMT) by comparing SG-GAN-2K and baseline approaches pairwise. We use 500 virtual-world images with size of 256×512 as input, and present pairs of adapted images generated by different methods to workers for A/B tests. For each image-image pair, we ask workers which image is more realistic than the other and record their answers. There are 123 workers participated in our A/B tests and the results are shown in Table 1. According to the statistics SG-GAN shows its superiority over all other approaches by a high margin. We attribute such superiority to clearer boundaries and smoother textures achieved by soft gradient-sensitive loss, and personalized texture rendering with the help of semantic-aware discriminator.

4.3 Ablation studies

Effectiveness of soft gradient-sensitive objective. To demonstrate the effectiveness of soft gradient-sensitive loss L_{grad}, we train a variant of SG-GAN without applying L_{grad} and compare it with SG-GAN-25K. Figure 5 shows an example by inspecting details through a 4X zoom. Compared with SG-GAN-25K, the variant without L_{grad} has coarse semantic boundaries and rough textures, which demonstrates soft gradient-sensitive loss can help generate adapted images with clearer semantic boundaries and smoother textures.

Effectiveness of semantic-aware discriminator. We use a variant of SG-GAN without applying semantic-aware discriminator (*SD*) and compare it with SG-GAN-25K to study the effectiveness of *SD*. As shown in Figure 6, comparing (g) and (h), the variant without *SD* lacks for details, e.g., the color of traffic light, and generates coarser textures, e.g., the sky. The difference maps, i.e., (b), (c), (d) in Figure 6, further reveal that semantic-aware discriminator leads to personalized texture rendering for each distinct region with specific semantic meaning.

The effect of virtual training image size. Figure 4 compares variants of SG-GAN that use distinct numbers of virtual-world images for training. Generally, SG-GAN-25K generates clearer details than SG-GAN-2K for some images. Further A/B tests between them in Table 1 show SG-GAN-25K is slightly better than SG-GAN-2K because of using more training data. Both qualitative and quantitative comparisons indicate more data could help, however, the improved performance may be only notable if dataset difference is in orders of magnitude.

4.4 Application on semantic segmentation

To further demonstrate the scene adaption quality of SG-GAN, we conduce comparisons on the downstream semantic segmentation task on Cityscapes validation set [3] by adapting from GTA-V dataset [24], similar to [13]. The idea is to train semantic segmentation model merely based on adapted virtual-world data, i.e., 24966 images of GTA-V dataset [24], and evaluate model's performance on real-world data, i.e., Cityscapes validation set [3]. For the semantic segmentation model we use the architecture proposed by [30] and exactly follow its training procedure, which shows impressive results on Cityscapes dataset. Table 2 shows the results. The baseline method is the version that trains semantic segmentation model directly on original virtual-world data and groundtruth pairs.

We first compare SG-GAN with CycleGAN [33]. The substantially higher semantic segmentation performance by SG-GAN shows its ability to yield adapted images closer to real-world data distribution. Figure 7 illustrates the visual comparison between SG-GAN and baseline to further show how SG-GAN helps improve segmentation. We further compare our approach with a hidden feature representation based adaption method proposed by [13], and SG-GAN achieves a high performance margin. These evaluations on semantic segmentation again confirm SG-GAN's ability to adapt high quality images, benefiting from preserving consistent semantic information and rendering personalized texture closer to real-world via soft gradient-sensitive objective and semantic discriminator.

5 Conclusion

In this work, we propose a novel SG-GAN for virtual-to-real urban scene adaption with the good property of retaining critical semantic information. SG-GAN employs a new soft gradient-sensitive loss to confine clear semantic boundaries and smooth adapted texture, and a semantic-aware discriminator to personalize texture rendering. We conduct extensive experiments to compare SG-GAN with other state-of-the-art domain adaption approaches both qualitatively and quantitatively, which all demonstrate the superiority of SG-GAN. Further experiments on the downstream semantic segmentation confirm the effectiveness of SG-GAN in virtual-to-real urban scene adaption. In future, we plan to apply our model on Playing-for-Benchmarks [25] dataset, which has an order of magnitude more annotated data from virtual-world for further boosting adaption performance.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow:

Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[4] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[5] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[6] LA Gatys, AS Ecker, and M Bethge. A neural algorithm of artistic style. *Nature Communications*, 2015.

[7] Leon A Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

[8] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *NIPS*, 2014.

[12] Judy Hoffman, Deepak Pathak, Eric Tzeng, Jonathan Long, Sergio Guadarrama, Trevor Darrell, and Kate Saenko. Large scale visual recognition through adaptation using joint representation and multiple instance learning. *JMLR*, 2016.

[13] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[16] M. Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, 2017.

[17] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[18] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. *arXiv preprint arXiv:1703.07022*, 2017.

[19] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, 2017.

[20] Xiaodan Liang, Hao Zhang, and Eric P Xing. Generative semantic manipulation with contrasting gan. *arXiv preprint arXiv:1708.00315*, 2017.

[21] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017.

[23] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv preprint arXiv:1611.04076*, 2016.

[24] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[25] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *ICCV*, 2017.

[26] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.

[27] Irvin Sobel. An isotropic 3× 3 image gradient operator. *Machine vision for three-dimensional scenes*, 1990.

[28] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[29] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[30] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016.

[31] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. *arXiv preprint arXiv:1709.07592*, 2017.

[32] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.