

Global and Local Texture Randomization for Synthetic-to-Real Semantic Segmentation

Duo Peng^{ID}, Yinjie Lei^{ID}, Member, IEEE, Lingqiao Liu^{ID}, Pingping Zhang^{ID}, and Jun Liu^{ID}, Member, IEEE

Abstract—Semantic segmentation is a crucial image understanding task, where each pixel of image is categorized into a corresponding label. Since the pixel-wise labeling for ground-truth is tedious and labor intensive, in practical applications, many works exploit the synthetic images to train the model for real-word image semantic segmentation, i.e., Synthetic-to-Real Semantic Segmentation (SRSS). However, Deep Convolutional Neural Networks (CNNs) trained on the source synthetic data may not generalize well to the target real-world data. To address this problem, there has been rapidly growing interest in Domain Adaption technique to mitigate the domain mismatch between the synthetic and real-world images. Besides, Domain Generalization technique is another solution to handle SRSS. In contrast to Domain Adaption, Domain Generalization seeks to address SRSS without accessing any data of the target domain during training. In this work, we propose two simple yet effective texture randomization mechanisms, Global Texture Randomization (GTR) and Local Texture Randomization (LTR), for Domain Generalization based SRSS. GTR is proposed to randomize the texture of source images into diverse unreal texture styles. It aims to alleviate the reliance of the network on texture while promoting the learning of the domain-invariant cues. In addition, we find the texture difference is not always occurred in entire image and may only appear in some local areas. Therefore, we further propose a LTR mechanism to generate diverse local regions for partially stylizing the source images. Finally, we implement a regularization of Consistency between GTR and LTR (CGL) aiming to harmonize the two proposed mechanisms during training. Extensive experiments on five publicly available datasets (i.e., GTA5, SYNTHIA, Cityscapes, BDDS and Mapillary) with various SRSS settings (i.e., GTA5/SYNTHIA to Cityscapes/BDDS/Mapillary) demonstrate that the proposed method is superior to the state-of-the-art methods for domain generalization based SRSS.

Index Terms—Synthetic-to-real semantic segmentation, domain generalization, texture randomization, consistency regularization.

Manuscript received August 14, 2020; revised March 29, 2021 and June 2, 2021; accepted July 6, 2021. Date of publication July 16, 2021; date of current version July 22, 2021. This work was supported in part by the Key Research and Development Program of Sichuan Province under Grant 2019YFG0409 and in part by the Fundamental Research Funds for the Central Universities under Grant DUT20RC(3)083. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jiaying Liu. (*Corresponding author: Yinjie Lei*)

Duo Peng and Yinjie Lei are with the College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China (e-mail: duo_peng@stu.scu.edu.cn; yinjie@scu.edu.cn).

Lingqiao Liu is with the School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia (e-mail: lingqiao.liu@adelaide.edu.au).

Pingping Zhang is with the School of Artificial Intelligence, Dalian University of Technology, Dalian 116024, China (e-mail: zhpp@dlut.edu.cn).

Jun Liu is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: jun_liu@sutd.edu.sg).

Digital Object Identifier 10.1109/TIP.2021.3096334

I. INTRODUCTION

SEMANTIC segmentation is a crucial computer vision task for various practical applications, such as robotic navigation and self-driving system [1]–[4]. For semantic segmentation, pixel-wise labeling is expensive and time consuming. Therefore, exploring economical methods for automatically generating annotations is appealing, i.e., using synthetic data whose ground-truth is straightforwardly provided [5], [6]. To this end, extensive works have been conducted for **Synthetic-to-Real Semantic Segmentation** (SRSS). However, the learned model is often challenged in practical scenarios where exists large difference between source domain (synthetic images) and target domain (real-world images). To solve this problem, **Domain Adaptation** (DA) [7]–[9] has recently been a research hot-spot. As shown in Fig. 1a, it aims to narrow the domain gap by introducing the unlabeled data from the target domain during training. However, data from the target domain is not always accessible beforehand in practical scenarios. Take autonomous driving for example, it is almost impossible to know the vehicle will under which situation in advance (e.g., the unknown weather, unknown city and unknown season). Based on the observations above, another growing trend of SRSS is **Domain Generalization** (DG) [10]–[13]. As shown in Fig. 1b, DG is aimed at training without access to any target domain data.

In this work, we propose two texture randomization mechanisms, one consistency regularization and a painting selection strategy to handle the Domain Generalization based SRSS. More specifically, such two mechanisms are proposed to generate augmented images by merging various texture appearances from paintings, i.e., **Global Texture Randomization (GTR)** and **Local Texture Randomization (LTR)**. Next, we train a deep CNN model for semantic segmentation with **Consistency between GTR and LTR (CGL)** to harmonize the proposed two randomization mechanisms. Besides, a strategy namely **Texture Complexity based Painting Selection (TCPS)** is proposed to ensure the selected paintings are reliable enough for the above texture randomization mechanisms. The next will be followed by a motivation on GTR, LTR, CGL and TCPS.

A. Global Texture Randomization (GTR)

Geirhos *et al.* [14] has proved that CNNs are strongly biased towards the texture of training images. However, the texture appearances between the synthetic and real-world

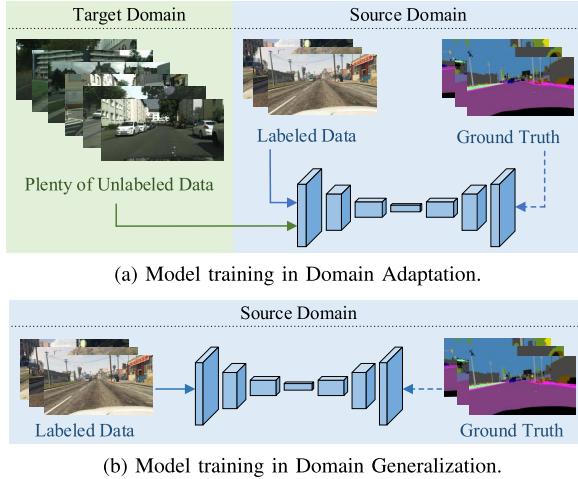


Fig. 1. Comparison of the model training between domain adaptation and domain generalization. Subfigure a: Domain adaptation, which assumes that the target domain is known beforehand and utilizes the unlabeled data of target domain during training. Subfigure b: Domain generalization only uses the labeled images from source domain for training.

image are significantly different (see Fig. 3), resulting in unsatisfactory SRSS performances. Existing methods [7], [15] take advantages of such network texture bias by transferring the style of source images to that of **real images**. As shown in Fig. 2, the real image can transfer the raw image into the domain of real-world scenario where the texture is regular and human-identifiable. However, when data from target domain is not accessible, these methods aimlessly seek various real images with diverse styles. This leads to inconvenience due to styles of real images should be manually selected in order to cover various possible scenarios that may occur in the target domain.

On the contrary, our approach leverages **unreal paintings** to handle style transfer. In the training phase, GTR is utilized to introduce more unreal textures which are random and irregular (shown in Fig. 5), making the difficulty of learning on image texture. Therefore, the network turns to learn more domain-invariant cues (e.g. shape and spatial layouts). As shown in Fig. 2, compared to transfer with real image, using unreal paintings can make the transferred texture irregular and indistinct. Since the unreal texture is utilized to promote domain invariant learning instead of introducing new domains, the style of painting does not matter. The introduction of GTR-generated images essentially alleviate the network heavy reliance on image textures, making the model more robust to the global texture changes caused by the synthetic-to-real shift. To the best of our knowledge, this is one of the early works applying unreal textures to model training for SRSS.

B. Local Texture Randomization (LTR)

GTR tends to make the network completely ignore the texture information, which is mixed blessing: it indeed strengthens the model robustness to texture shift but also prevent the network from utilizing useful texture cues for segmentation. In fact, in most cases, only some local areas of the image

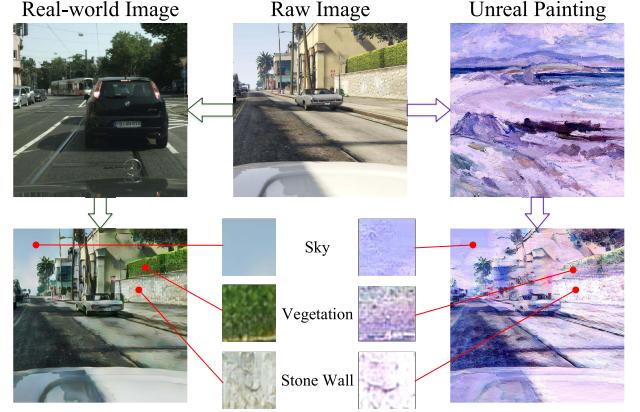


Fig. 2. Comparison between style transfer with real image and unreal painting.

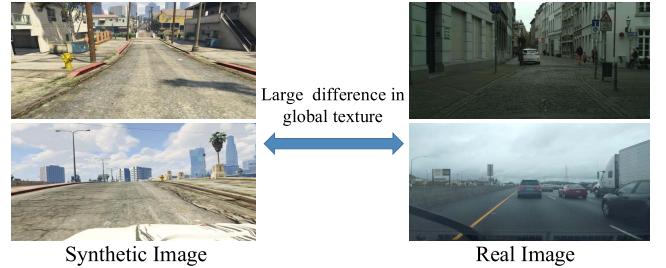


Fig. 3. Examples of images with large texture difference. It can be seen that the textures between the synthetic image and real image are globally different.

are with texture differences. As shown in Fig. 4, only the areas masked in red color contain large texture difference, while other areas are with small texture discrepancy. Thus, we propose a Local Texture Randomization (LTR) to address this situation. More specifically, LTR mechanism is proposed to generate images with local randomized texture for training. Since such local areas are usually with arbitrary shapes, LTR generates the local texture randomization areas with random boundaries (shown in Fig. 6). Therefore, the usage of LTR enables CNNs to achieve a favorable performance under various local texture-different scenarios.

C. Consistency Between GTR and LTR (CGL)

In the proposed framework, both GTR and LTR are utilized for source image texture randomization. Nevertheless, source images simultaneously using both global and local texture randomization mechanisms may drive the network toward distinct learning directions and fail to converge. To solve this problem, we propose a CGL mechanism to harmonize the proposed global and local texture randomization mechanisms during the training phase.

D. Texture Complexity Based Painting Selection (TCPS)

Since unreal textures are required to randomize source images, we utilize the paintings from dataset ‘Painter by Numbers’.¹ However, we observe that paintings with different

¹<https://www.kaggle.com/c/painter-by-numbers/>

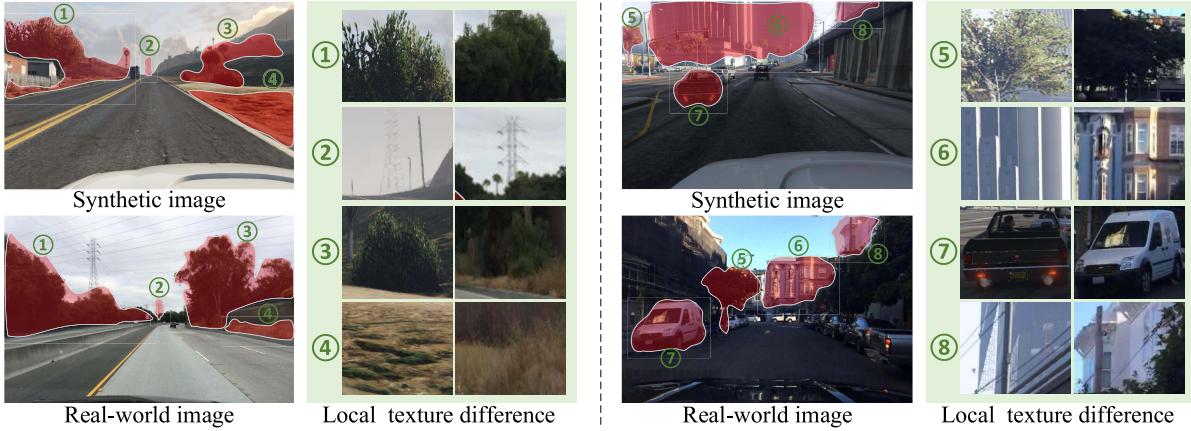


Fig. 4. Examples of images with local texture difference. It can be observed that only local areas between synthetic and real images are texture-different, especially for vegetation, terrain and building.

texture complexities can significantly affect the model performance, while different style paintings with a same texture complexity lead to similar performances. In other words, the SRSS performance is sensitive to texture complexity instead of style of unreal paintings. Therefore it is necessary to give a measurement method for texture complexity so as to experimentally find what extent of texture complexity is suitable to SRSS. To this end, we propose a Texture Complexity based Painting Selection (TCPS) strategy to measure texture complexity and select reliable paintings.

The contributions of this paper are summarized as follows:

- We propose to utilize GTR mechanism to randomly replace the texture of source images with unreal textures, aiming to alleviate the network's reliance on texture.
- We propose a LTR mechanism to further cover the cases in which only some local areas contain large texture difference between the synthetic and real images.
- Moreover, we propose a consistency regularization between GTR and LTR to harmonize the proposed global and local texture randomization mechanisms during training.
- Besides, we propose a painting selection strategy based on measurement of texture complexity to select reliable paintings for texture randomization.
- The proposed method is comprehensively evaluated under various SRSS settings (i.e., GTA5/SYNTHIA to Cityscapes/BDDS/Mapillary), obtaining superior segmentation performance over other state-of-the-arts.

II. RELATED WORKS

A. Semantic Segmentation

Semantic segmentation is a crucial computer vision task and has attracted growing research interest as its various real-world applications. Before the advent of deep learning, traditional optimization techniques such as graph cuts [16], belief propagation [17], message passing [18], and random walks [19] have been widely used in semantic segmentation. For instance, [20] proposes a common framework which unifies graph cuts and random walks. References [21] proposes a sub-Markov random walk algorithm to solve the problem of objects with thin and elongated parts. However,

the energy formulations for describing natural images are complicated. Therefore higher-order energy functions [22] have been attracted more attention. References [23] proposes a framework of maximizing quadratic submodular energy with a more general knapsack constraint. Besides, an energy minimization method for general explicit higher-order energy functions is proposed [24]. It is competent in image segmentation with the appearance entropy.

As for deep learning based methods, the most frequently-used framework is Fully Convolutional Network (FCN) [2]. It is capable of constructing an end-to-end mapping between the input RGB image and the semantic pixels. Afterwards, dilated convolution [25] has been proposed to modify the original FCN. Meanwhile, Markov and conditional random fields [26]–[28] are exploited to make up for the loss of object details. Afterwards, Global Convolutional Network [29] is presented to use large kernels to handle semantic segmentation. References [30] uses semantic enhancement module with feature attention for fusing of shallow features and deep features. Besides, Kernel-Sharing Atrous Convolution (KSAC) [31] is proposed to share different parallel layers.

B. Domain Adaptation

It is known that, collecting a plenty of annotated training image data for each city of interest would be computationally expensive. To address this issue, an enormous amount of DA techniques are proposed. First of all, Richter *et al.* [6] gave an exploration on extracting images from modern computer games with annotations generated by engine, inspiring a series of further studies. Most existing works fall into three categories: feature-level adaptation, pixel-level adaptation and both. (1) Feature-level adaptation aims to align features by either directly minimizing the feature distance [32]–[35] or implicitly narrowing the gap between source and target distributions with adversarial learning [8], [36]–[41]. (2) Pixel-level adaptation attempts to remove representation differences by stylizing source images to realistic target images [15], [42]–[44]. (3) The rest of methods take both levels described above into consideration [7], [45]. Even in a same adaptation level, methods varies at different alignment

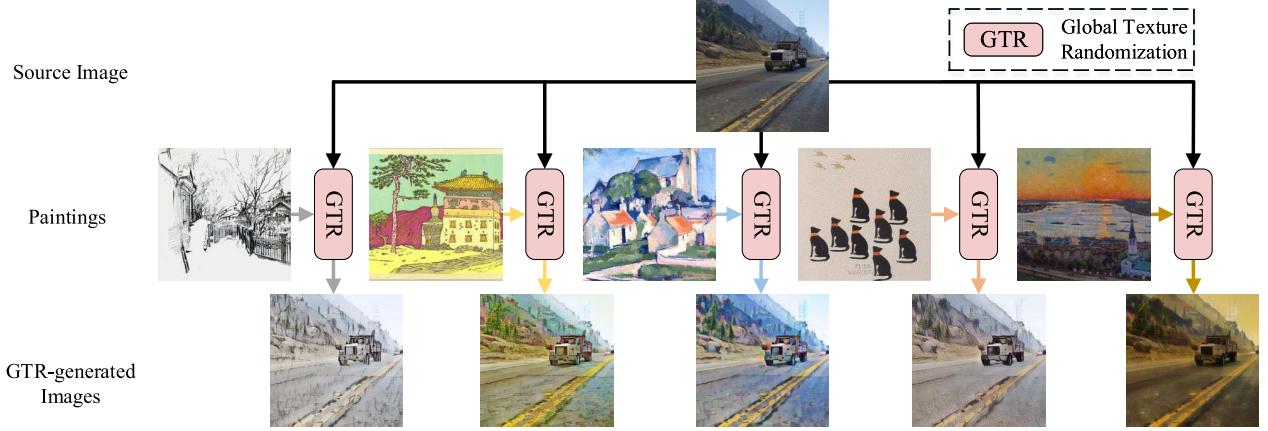


Fig. 5. Global texture randomization. Top: a cropped source image. Mid: texture images composed of paintings with various styles. Bottom: GTR-generated images with the same content information of source image and identical textures of paintings.

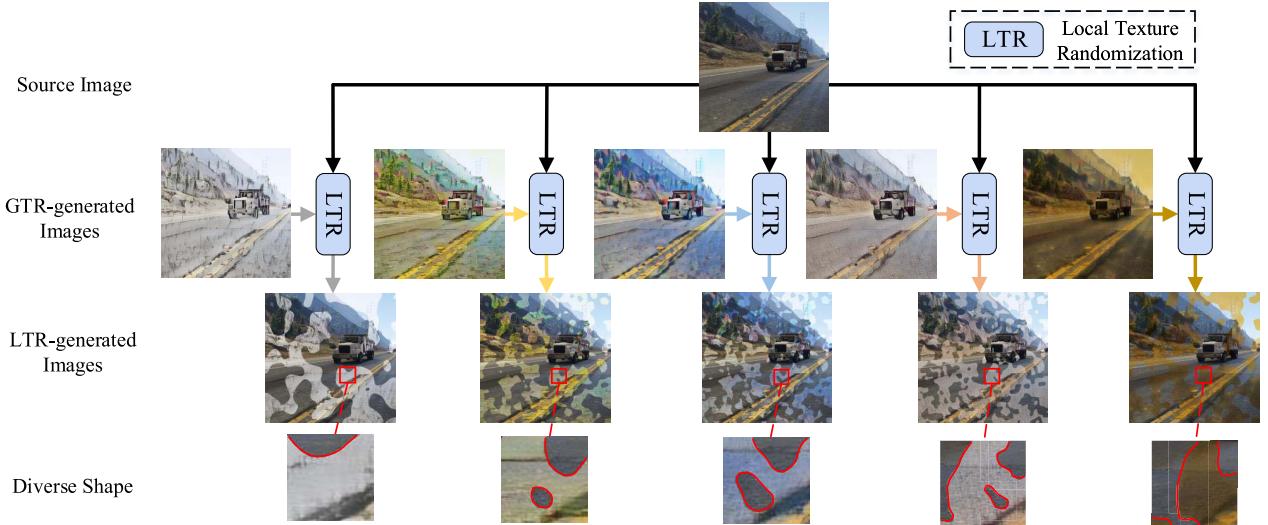


Fig. 6. Local texture randomization. Aiming to obtain LTR-generated images (shown in the third row) which contains randomization regions with diverse shapes (shown in the last row), we mix the source image (shown in the first row) and each GTR-generated image (shown in the second row) respectively.

perspectives: class-wise alignment [36], [39], [46], patch-wise alignment [41], layer-wise alignment [33], [35] and label-wise alignment [37], [44], etc.

C. Domain Generalization

Unlike Domain Adaptation, Domain Generalization purely utilizes the labeled images in source domain [47], [48]. References [49] explores latent domains among training and testing data. References [50] presents a framework based on Maximum Mean Discrepancy (MMD) which measures the generalized latent feature representation across different domains. References [51] develops a CNN model for an end-to-end Domain Generalization learning. References [10] utilizes the meta-learning strategy for the improvement of generalization performance. Although above methods are effective, the majority of them mainly focused on image classification. There are only a few DG studies addressing semantic segmentation [52]–[54], which can be broadly classified as: layer-normalized generalization and data-enhanced generalization. (1) Layer-normalized generalization propose to search a

reasonable combination of different whitening and standardization techniques in CNNs so as to improve generalization performance on unseen domains [53], [54]. (2) Data-enhanced generalization tends to randomize the synthetic images with various styles of real images, in order to effectively cover real-world target domain [52]. Although both [53] and [54] are layer normalized generalization, [54] further takes whitening techniques (BW [55] and IW [56]) into consideration compared with [53] that only utilizes standardization techniques (IN [57] and BN [58]). However, training with various unreal features has been a largely under explored domain. In this paper, we propose to handle DG with randomized unreal images which implemented by Data Enhancement.

D. Data Enhancement

In this paper, we mainly provide a Data Enhancement method which resorts to Regularization Techniques and Data Augmentation.

1) *Data Augmentation*: Data Augmentation is the process of supplementing an extra dataset with similar data created

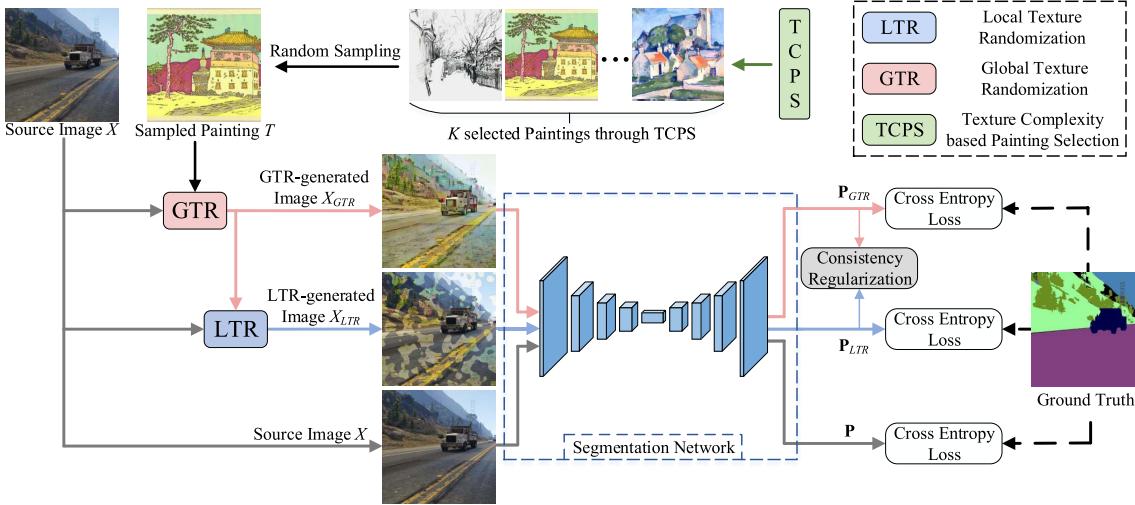


Fig. 7. The overall architecture of our framework. After TCPS, K paintings are obtained by random sampling. For each iteration, we randomly select one painting from K paintings, namely T . With the help of the GTR/LTR mechanism, X_{GTR}/X_{LTR} generated. There are 3 streams for network training. Besides the cross entropy loss on each stream, a consistency regularization is applied between GTR stream and LTR stream to harmonize the learning.

from the raw information which is crucial in deep learning. When dealing with images, it often handles the image by rotating, cropping, blurring, adding noise and other operations to existing images, enabling the network to generalize better [59]–[62]. References [63] automatically merges multiple images from the same class for data augmentation. A Bayesian based approach is proposed in [64] to augment the training set. References [65] applies simple transformations such as interpolating and extrapolating to perform the transformation in a learned feature space. In [66], the misclassified examples are utilized for data augmentation.

2) *Regularization Techniques*: To generate randomized images, Regularization Techniques are also utilized. Hence, some common Regularization Techniques are introduced in this part. First, the MixUp regularization [67] improves the performance of network training by randomly interpolating samples from training set. Then, the Cutout regularization [68] augments the training data by shielding a rectangular region in image. Recently, [69] presents CutMix regularization which combines both MixUp and Cutout. Due to the addition of a stronger perturbation, CutMix improves model performance and outperforms the other two methods (i.e., MixUp and Cutout).

III. METHODOLOGY

A. Overview of the Proposed Framework

The overall architecture of our framework is shown in Fig. 7. First, we utilize TCPS to execute pre-selection towards dataset ‘Painter by Numbers’ and obtain a subset of dataset. Then K paintings are given by random sampling from subset. In this paper, the hyper-parameter K is set to 15. Since our goal is to randomize the texture of source image during training, one painting T is randomly selected from 15 paintings in each iteration. Formally, we consider a source image $\{X\} \in \mathbb{R}^{H \times W \times 3}$ along with a selected painting T . The image X and painting T are fed forward into GTR module

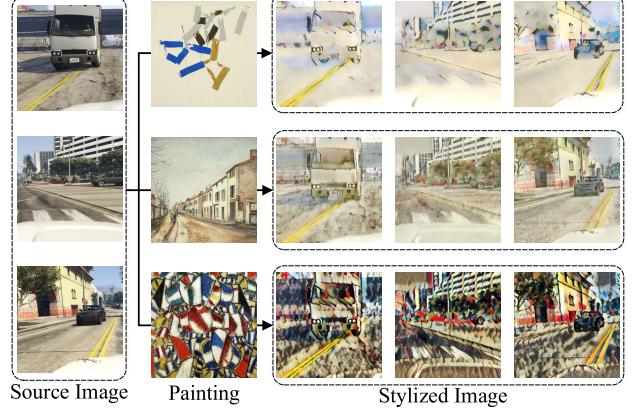


Fig. 8. Texture randomization results using paintings with different texture complexities. The texture complexity of painting increases from the top to bottom.

to generate a stylized image X_{GTR} . Next, we simultaneously send X and X_{GTR} into LTR to generate a local stylized image X_{LTR} . Afterwards, X , X_{GTR} and X_{LTR} are fed into the segmentation network to obtain segmentation outputs P , P_{GTR} and P_{LTR} respectively. For a better convergence, consistency is applied between GTR and LTR to make segmentation predictions (i.e., P_{GTR} and P_{LTR}) close to each other. Finally, each stream of segmentation losses is computed to update the network’s weights. During testing, we remove these modules and only retain the segmentation network.

B. Texture Complexity Based Painting Selection (TCPS)

From Fig. 8, we can see that the stylized image performs unfavorably when texture complexity of the painting becomes quite high (i.e., the third row) or low (i.e., the first row). The key to ensure the generalization performance is to select paintings which can effectively replace image texture while preserving the shape of objects. In light of this, we propose



Fig. 9. Examples of smooth-unsound area separation from dataset “Painter by numbers”. The first column shows the raw paintings. The last two columns are the separated smooth area and the corresponding unsound area.

a Texture Complexity measurement to select paintings from dataset “Painter by Numbers”.

According to the gradient representation of each pixel, we divide an image into two parts: (1) smooth area, where pixel values change slightly or have no change; (2) unsound area, where pixel values change significantly. Inspired by Structure Tensor [70], we separate the two areas by calculation of gradient representation for each pixel. Assume $Grad(i, j)$ denotes the gradient representation of pixel at row i and column j .

The gradient representation $Grad(i, j)$ can reflect the extent of change in pixel value and hence indicates whether the area of pixel is smooth or not. Pixels with little gradient representation are formed into the smooth area, such as sky, wall and background. On the contrary, pixels with large gradient representation make up the unsound area which has a distinct texture expression. To quantitatively determine which area the pixel belongs to, we set a threshold ε manually. When meeting the condition: $Grad(i, j) < \varepsilon$, the pixel is allocated to the smooth area, or else it belongs to the unsound area. In this paper, we set ε to 20. Fig. 9 shows some examples of smooth-unsound area separation by using our method. Qualitative results demonstrate the effectiveness of our method.

As textures of an image almost gather in unsound area, the texture complexity can be represented by the proportion of unsound area in the image. It can be formulated as following:

$$\mathcal{T}_{exc} = \frac{\sum_i \sum_j \text{Bool}(I(i, j) \in A_{usm})}{N} \quad (1)$$

where $\text{Bool}(\cdot)$ is the judgment function, if the condition is true it returns 1, or else it returns 0; A_{usm} denotes the unsound area; N is the total number of pixels. In essence, the texture complexity is the proportion of unsound area in the image. By calculating the texture complexity, we can determine if

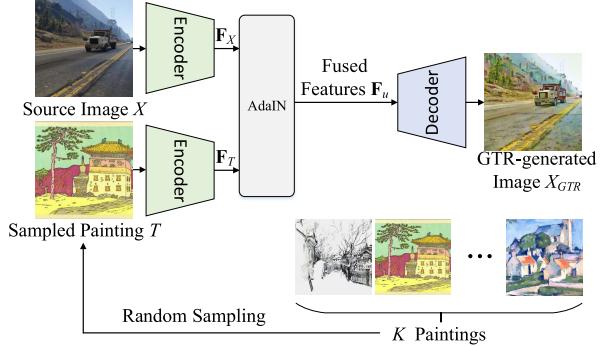


Fig. 10. The detailed architecture of the GTR module. The major structure of GTR module is a pre-trained style translation network [71]. Aiming to replace the texture of the raw image X to that of the painting T , an AdaIN layer [71] is utilized to fuse the features extracted by the encoder. Then, the fused feature map \mathbf{F}_u is fed forward to the decoder to obtain the GTR image X_{GTR} .

this painting is competent to carry out texture randomization. Based on the ablation study in Section V-B, we find the painting with $\mathcal{T}_{exc} \in [0.55, 0.65]$ can achieve an pleasurable and stable texture randomization performance.

C. Global Texture Randomization (GTR)

GTR aims to strip the texture of the source image and replace with the style of the painting. Note that in each iteration, we implement random sampling on K paintings to acquire the painting T . We then use the pre-trained style translation network [71] to transfer the texture of the selected painting to the source image X . More specifically, As shown in Fig. 10, the raw image X and the painting T are first fed into a pre-trained Siamese encoder for feature extraction. The feature extraction can be formulated as following:

$$\mathbf{F}_X = f(X, \theta^{en}) \quad (2)$$

$$\mathbf{F}_T = f(T, \theta^{en}), \quad (3)$$

where \mathbf{F}_X and \mathbf{F}_T denote the extracted feature maps from source image X and painting T , θ^{en} denotes the weights of the encoder network $f(\cdot, \cdot)$.

Then an Adaptive Instance Normalization (AdaIN) [71] is utilized to perform feature fusion. Particularly, AdaIN treats the mean and variance of the extracted feature maps (i.e., \mathbf{F}_X and \mathbf{F}_T) as affine parameters. The feature fusion implemented by AdaIN can be written as following:

$$\mathbf{F}_u = \sigma(\mathbf{F}_T) \cdot \left(\frac{\mathbf{F}_X - \mu(\mathbf{F}_X)}{\sigma(\mathbf{F}_X)} \right) + \mu(\mathbf{F}_T), \quad (4)$$

where \mathbf{F}_u is the fused feature map, $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and variance values which are spatially computed across the entire feature maps.

Finally, a pre-trained decoder inverts the transferred feature map \mathbf{F}_u to the GTR image X_{GTR} . Formally, the GTR-generated image X_{GTR} can be calculated by:

$$X_{GTR} = f(\mathbf{F}_u, \theta^{de}), \quad (5)$$

where θ^{de} denotes the weights of the decoder network.

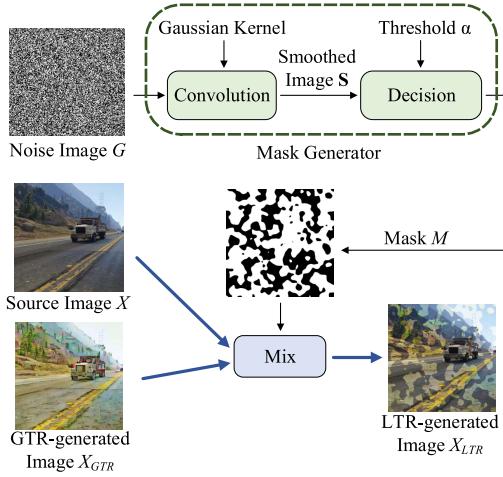


Fig. 11. Illustration of the proposed LTR module. To generate the mask, a well-designed mask generator [72] is applied to a noise image.

After the process of GTR module, the texture variety of training set is augmented, leading to robustness of global texture change. Despite such effectiveness, the model with GTR ignores the cases where only local texture difference exists between synthetic and real images. We improve it by the following Local Texture Randomization.

D. Local Texture Randomization (LTR)

LTR is proposed to randomly generate images with local randomized textures. For LTR module, the raw image X and GTR-generated image X_{GTR} are fed into LTR module to generate the output image X_{LTR} . In each iteration, LTR generates a mask M with a random shape to mix the source image and GTR-generated image so that the masks among different iterations are diverse. As shown in Fig. 11, we utilize the mask generator [72] to create mask M . For more details, we firstly generate an image with Gaussian noise which follows a normal distribution $\mathcal{N}(0, 1)$, named as G . Then, the noise image G is smoothed with a Gaussian kernel:

$$\mathbf{S} = f^{3 \times 3}(G, \gamma), \quad (6)$$

where \mathbf{S} denotes the smoothed image, $f^{3 \times 3}(\cdot)$ denotes the convolution with 3×3 Gaussian kernel and γ is the standard deviation of Gaussian kernel. As for γ , we use a nonlinear calculation on variable parameter λ to obtain it:

$$\gamma = \exp(\lg(\lambda)). \quad (7)$$

In each training iteration, we obtain λ from the uniform distribution $[\lambda_{min}, \lambda_{max}]$ by random selection. After ablation study in Section V-B, we respectively set λ_{min} and λ_{max} to 4 and 16 in this paper. In particular, the random determination of λ leads to the variety of mask. Fig. 12 shows some generated masks with varying values of λ .

At the stage of generating mask, a threshold α is used to decide whether each pixel is set to 0 (black) or 1 (white). For a mask which contains white region with a proportion of p , we can obtain the threshold α through ranking all pixel values and choosing the last value of former p elements.

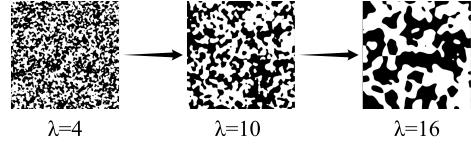


Fig. 12. Example of masks varying λ from 4 to 16 with resolution 640×640.

However, ranking algorithm is computationally expensive, leading to slow training speed. To alleviate this problem, previous study [72] has shown that this goal can be achieved through the error function [73] without ranking pixel values. Therefore, given a proportion p , we can directly compute the threshold α as following:

$$\alpha = \text{erf}^{-1}(2p - 1) \cdot \sqrt{2}\sigma_S + \mu_S, \quad (8)$$

where $\text{erf}^{-1}(\cdot)$ is the inverse function of error function $\text{erf}(\cdot)$ [73], μ_S and σ_S are the mean value and standard deviation of the smoothed image S respectively. Particularly, p is the hyper parameter to decide the proportion of white region. After obtaining the threshold α , we turn the smoothed image S into mask M . The i -th element M_i of the mask M can be calculated by:

$$M_i = \begin{cases} 0 & E_i < \alpha \\ 1 & E_i \geq \alpha \end{cases}, \quad (9)$$

where E_i denotes the i -th element in S .

Finally, we mix the raw image X and GTR-generated image X_{GTR} with mask M . We define the mix function as $\text{Mix}(X, X_{GTR}, M)$. Specifically, the LTR-generated image X_{LTR} can be computed as:

$$X_{LTR} = \text{Mix}(X, X_{GTR}, M) = (1-M) \odot X + M \odot X_{GTR}, \quad (10)$$

where \odot denotes pixel-wise multiplication. With this step, X_{LTR} is generated to further improve the generalization ability.

E. Consistency Between GTR and LTR (CGL)

Simply training the network with GTR images and LTR images has two weaknesses: (a) different randomization methods will fail to converge for network training; (b) even if the network converge well, the network may be trained towards completely different directions resulting in unstable performance. To this end, we impose the consistency loss to harmonizes the two randomization learning methods. L_1 distance is utilized for the calculation of consistency loss. The formulation of proposed CGL can be written as:

$$\mathcal{L}_{con}(X_{GTR}, X_{LTR}) = \mathcal{L}_1(f(X_{GTR}, \theta), f(X_{LTR}, \theta)), \quad (11)$$

where $\mathcal{L}_1(\cdot, \cdot)$ denotes the L_1 distance, θ is the weights of the whole segmentation network $f(\cdot, \cdot)$. Note that we extract the feature maps from the last convolutional layer before softmax to calculate the L_1 distance.

F. Network Training

Given a set of training samples $\mathcal{D} = \{X, X_{GTR}, X_{LTR}, Y\}$, where X , X_{GTR} and X_{LTR} are the input raw image, GTR-generated image and LTR-generated images, Y is the ground-truth. Considering both GTR and LTR are utilized in training, we define the following segmentation loss:

$$\mathcal{L}_{seg} = \mathcal{L}_w(X, Y) + \mathcal{L}_w(X_{GTR}, Y) + \mathcal{L}_w(X_{LTR}, Y), \quad (12)$$

where $\mathcal{L}_w(X, Y)$, $\mathcal{L}_w(X_{GTR}, Y)$ and $\mathcal{L}_w(X_{LTR}, Y)$ is the weighted cross entropy loss of raw image, GTR-generated image and LTR-generated image, respectively.

Considering the consistency regularization described in Section III-E, our final training loss can be rewritten as:

$$\mathcal{L} = \arg \min (1 - \beta) \mathcal{L}_{seg} + \beta \cdot \mathcal{L}_{con}(X_{GTR}, X_{LTR}), \quad (13)$$

where β is the hyper-parameter which trades off the quality of segmentation loss (\mathcal{L}_{seg}) with consistency loss (\mathcal{L}_{con}). In this paper, β is set to 1e-5. As for testing, the image only pass through the segmentation network without the process of GTR, LTR and CGL.

IV. EXPERIMENTAL SETUPS

A. Datasets Description

1) *Source Domain Datasets*: *GTA5* is a synthetic image dataset collected in a computer game with pixel-wise semantic labels. It contains 24966 images with resolution of 1914×1052 . It includes 19 classes which are compatible with most semantic segmentation datasets of outdoor scenes.

SYNTHIA is a large synthetic dataset with pixel-level semantic annotations. The subset *SYNTHIA-RANDCITYSCAPES* is used in our experiments which contains 9400 images. It consists of 16 categories with high resolution of 1280×760 .

2) *Target Domain Datasets*: *Cityscapes* contains vehicle-centric urban street images taken from some European cities. There are 5000 images with pixel-wise annotations. The images with size of 2048×1024 are labeled into 19 classes.

BDDS contains thousands of real-world dashcam video frames with accurate pixel-wise annotations. It contains 34 labeled categories with resolution of 1280×720 . The training, validation, and test sets contain 7000, 1000 and 2000 images, respectively.

Mapillary contains street-level images with diverse resolutions. The annotations contain 66 object classes, but only 19 and 16 classes that overlap with *GTA5* and *SYNTHIA* respectively, are used in our experiments. It has a training set with 18000 images and a validation set with 2000 images.

3) *SRSS Experimental Settings*: ① *GTA5* to *Cityscapes*, ② *GTA5* to *BDDS*, ③ *GTA5* to *Mapillary*, ④ *SYNTHIA* to *Cityscapes*, ⑤ *SYNTHIA* to *BDDS* and ⑥ *SYNTHIA* to *Mapillary*.

B. Implementation Details

For fair comparison with other Domain Generalization methods, we use VGG-16 [74], ResNet-50 and ResNet-101 [75] as the segmentation network respectively. The source

TABLE I

THE COMPARISON RESULTS WITH DRPC. THIS TABLE SHOWS THE GENERALIZATION RESULTS FOR SRSS SETTINGS FROM GTA5. THE BEST RESULTS OF EACH SETTING ARE MARKED IN BOLD

| Network | SRSS Setting | Methods | mIoU | mIoU ↑ |
|-------------------|--------------------|------------------|-------------|-------------|
| ResNet-101 | GTA5 to Cityscapes | Source Only [52] | 33.6 | 8.9 |
| | | DRPC [52] | 42.5 | |
| | GTA5 to BDDS | Source Only | 34.0 | |
| | | Ours | 43.7 | 9.7 |
| GTA5 to Mapillary | GTA5 to BDDS | Source Only [52] | 27.8 | |
| | | DRPC [52] | 38.7 | 10.9 |
| | GTA5 to Mapillary | Source Only | 28.1 | |
| | | Ours | 39.6 | 11.5 |
| GTA5 to Mapillary | GTA5 to Mapillary | Source Only [52] | 28.3 | |
| | | DRPC [52] | 38.1 | 9.8 |
| | | Source Only | 28.6 | |
| | | Ours | 39.1 | 10.5 |

codes and models are trained and evaluated on PyTorch toolbox [76] based on Python 3.6 platform.² For more details, all proposed models are implemented on 8 NVIDIA RTX 2080Ti GPUs and two E5-2620 CPUs.

1) *Evaluation Metric*: Following previous works, we utilize PASCAL VOC Intersection over Union (IoU) [77] as the evaluation metric for testing. And mIoU is the mean value of IoUs across all categories. There is a positive correlation between mIoU value and segmentation performance.

2) *Data Preprocessing*: For the training set in *GTA5*, we crop each panoramic image into two patches with a resolution of 1052×1052 . Then, we resize each cropped image into a resolution of 640×640 . For the training data in *SYNTHIA* dataset, we also crop each image into two patches with a resolution of 640×640 . Since downsampling of CNN, we restore the resolution of the output prediction by upsampling with Bi-linear Interpolation. Finally, as [78], [79] do, we further adopt random mirror and Gaussian blur to images before network training.

3) *Parameter Settings*: The model is initialized with the network parameters pre-trained on ImageNet [80] except those of the final classifier layer. In training period, we choose standard Stochastic Gradient Descent (SGD) [81] optimizer with a batch size of 2, a momentum of 0.9 and a weight decay of 5e-4. The learning rate is set to 1e-5 initially and follows the poly learning rate policy [82] with a poly power of 0.9. The total number of training iterations is set to 200000. We set the probability of triggering random mirror and Gaussian blur to 50%. In addition, the radius in Gaussian blur is randomly selected from the range of $[0, 1]$.

V. EXPERIMENTAL RESULTS

A. Comparison With State-of-Arts

In this subsection, extensive experiments are conducted to show the generalization performance of our framework. Since few work has been devoted to Domain Generalization, we compare our results with the only known three Domain Generalization methods: DPRC [52], IBN [53] and SW [54]. Note that DPRC, IBN and SW are based on different backbone networks, i.e., ResNet-101 (DPRC), ResNet-50 (IBN) and

²The source codes will be released upon the acceptance of this paper.

TABLE II

THE COMPARISON RESULTS WITH DRPC. THIS TABLE SHOWS THE GENERALIZATION RESULTS FOR SRSS SETTINGS FROM SYNTHIA. THE BEST RESULTS OF EACH SETTING ARE MARKED IN BOLD

| Network | SRSS Setting | Methods | mIoU | mIoU ↑ |
|-----------------------|------------------|------------------|-------------|------------|
| SYNTHIA to Cityscapes | Source Only [52] | Source Only | 29.7 | 7.9 |
| | | DRPC [52] | 37.6 | |
| | Ours | Source Only | 30.2 | 9.5 |
| | | Ours | 39.7 | |
| ResNet-101 | SYNTHIA to BDDS | Source Only [52] | 25.6 | 8.7 |
| | | DRPC [52] | 34.3 | |
| | Ours | Source Only | 25.9 | 9.5 |
| | | Ours | 35.3 | |
| SYNTHIA to Mapillary | Source Only [52] | Source Only | 28.7 | 5.4 |
| | | DRPC [52] | 34.1 | |
| | Ours | Source Only | 29.5 | 6.9 |
| | | Ours | 36.4 | |

TABLE III

QUANTITATIVE RESULTS OF THE COMPARISON WITH IBN. THE BEST RESULTS ARE MARKED IN BOLD

| Network | SRSS Setting | Methods | mIoU | mIoU ↑ |
|-----------|--------------------|------------------|-------------|------------|
| ResNet-50 | GTA5 to Cityscapes | Source Only [53] | 22.2 | |
| | | IBN [53] | 29.6 | 7.4 |
| | Ours | Source Only | 31.7 | |
| | | Ours | 38.6 | 6.9 |

TABLE IV

COMPARISON RESULTS OF OURS AND COMBINATION OF OURS AND IBN. THE BEST RESULTS ARE MARKED IN BOLD

| Network | SRSS Setting | Methods | mIoU |
|-----------|--------------------|---------------|-------------|
| ResNet-50 | GTA5 to Cityscapes | Source Only | 31.7 |
| | | Ours | 38.6 |
| | | Ours+IBN [53] | 39.5 |

VGG16 (SW). Therefore, we implement our method on three different backbone networks to make a fair comparison.

1) *Comparison With DRPC*: Following DRPC [52], we report the results on three different target domains (Cityscapes, BDDS and Mapillary) from two source domains (GTA5 and SYNTHIA). Specifically, we implement our framework in six experimental settings: “GTA5 to Cityscapes”, “GTA5 to BDDS”, “GTA5 to Mapillary”, “SYNTHIA to Cityscapes”, “SYNTHIA to BDDS” and “SYNTHIA to Mapillary”. DRPC [52] introduce many real-world domains in network training with a pyramid consistency regularization. As shown in Tab. I and Tab. II, we can see that our proposed framework achieves superior results in all settings. In addition, our framework also has a better mIoU gain in each experimental setting. This illustrates that our proposed framework is more capable of improving generalization performance.

2) *Comparison With IBN*: We follow the experimental setting (“GTA5 to Cityscapes”) in IBN [53] to evaluate our framework using backbone network ResNet-50. More specifically, the training set only contains the labeled data in GTA5. Since the ground-truth of testing set in Cityscapes is missing, we use the validation set as the testing set in our experiments. IBN utilizes Instance Normalization (IN) and Batch Normalization (BN) to capture more invariant information from appearance changes while preserving content related information. Our method also learns invariant information, but we devote to data randomization while IBN devotes to network structure design. From the comparison shown in Tab. III,

TABLE V

QUANTITATIVE RESULTS OF THE COMPARISON WITH SW. THE BEST RESULTS ARE OBTAINED BY IN-COLUMN COMPARISON AND MARKED IN BOLD

| Network | SRSS Setting | Methods | mIoU | mIoU ↑ |
|---------|--------------------|-------------|------|--------|
| VGG-16 | GTA5 to Cityscapes | Source Only | - | |
| | | SW [54] | 35.7 | |
| | | Ours | 31.4 | 5.8 |

it can be seen that although the mIoU gain of ours is slightly weaker than IBN, our framework achieves a mIoU of 38.6% which is significantly 9% better than IBN. In addition, our method is complementary with IBN. To demonstrate that, the additional experiment on model with both IBN and our method is conducted. The results are shown in Tab. IV. We can see that the model with both IBN and our method boosts the performance to 39.5%, which is 0.9% better than the model with only our method.

3) *Comparison With SW*: SW [54] is a modified model from IBN [53]. It provides an integration of several whitening techniques and standardization techniques. SW can adaptively set appropriate whitening or standardization statistics for a specific task. SW offers more possibilities for learning invariant information than IBN. Tab. V shows the performance of SW and our method under the experimental setting “GTA5 to Cityscapes”. With the same backbone network (VGG16), our framework achieves a mIoU of 37.2% and outperforms the SW model by 1.5%.

4) *Comparison With Domain Adaption Methods*: All methods above are conducted with Domain Generalization manner, where the network training has no access to target domain. Now we compare our results with previous state-of-the-art Domain Adaption methods which are target domain-accessible. In view of the most extensive Domain Adaption methods are based on VGG-16 and ResNet-101, we conduct experiments on the same backbone networks. Since most of previous works conducted domain adaption from GTA5/SYNTHIA to Cityscapes, we present the comparison results under the same SRSS settings in Tab. VI and Tab. VII, respectively. It can be seen that in each setting, our method is on par or even better than existing Domain Adaption methods.

B. Ablation Studies

In this subsection, we demonstrate the benefit of each design in our approach. Without loss of generality, we only exhibit the generalization results from GTA5 (G) to Cityscapes (C), BDDS (B) and Mapillary (M) with the backbone network ResNet-101. In Tab. VIII, we detail the mIoU improvement by progressively adding components: Global Texture Randomization (GTR), Local Texture Randomization (LTR) and Consistency between GTR and LTR (CGL). Besides, to demonstrate benefits of the proposed Texture Complexity based Painting Selection (TCPS), the same ablation experiments are conducted under different conditions (with or without TCPS).

1) *Effects of Global Texture Randomization (GTR)*: One can observe that compared to baseline (Tab. VIII a), the proposed GTR (Tab. VIII b) helps boost the performance from 34.0%

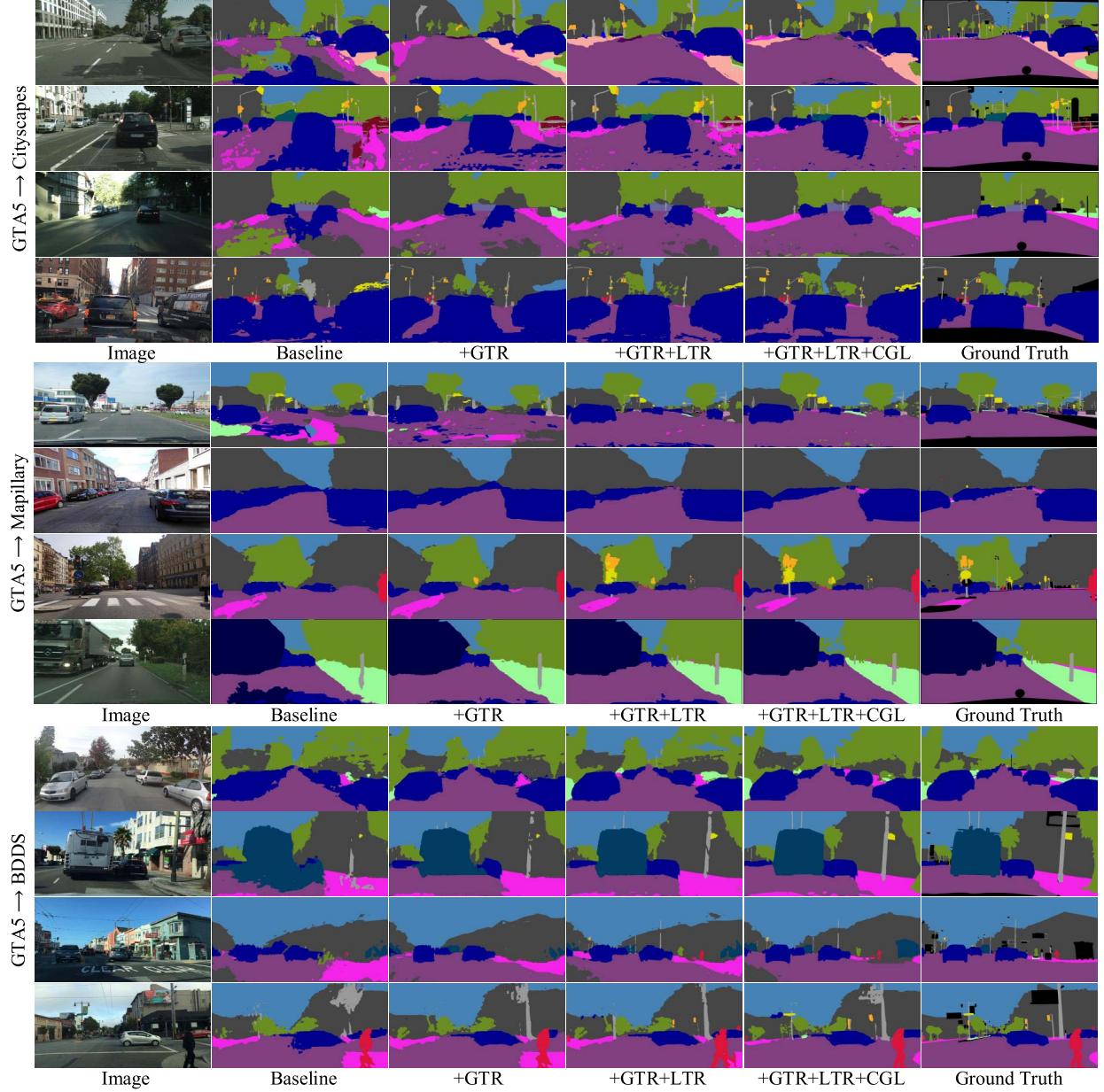


Fig. 13. Qualitative domain generalization based SRSS results from GTA5 to cityscapes, Mapillary and BDDS. It can be seen that the segmentation performance is improved by adding the proposed modules progressively. Note that all results are obtained under the condition of TCPS. (The figure is best viewed in color.)

to 39.2%, from 28.1% to 35.4% and from 28.6% to 32.1% on GTA5 (G)→Cityscapes (C), GTA5 (G)→BDDS (B) and GTA5 (G)→Mapillary (M). We can also see that the model with GTR can achieve 5.2% improvement in experimental setting G→C, 7.3% improvement on G→B and 3.5% improvement on G→M. The improvements of the GTR clearly demonstrate the benefit of utilizing diverse texture styles.

2) Effects of Local Texture Randomization (LTR): To evaluate the benefits of the proposed module LTR, we also re-implement our approach only with LTR (Tab. VIII c). In terms of mIoU, our framework with LTR achieves 41.8%/36.1%/33.7% on G→C/G→B/G→M. Meanwhile, compared to baseline (Tab. VIII a), model with LTR can greatly boost the results by 7.8%/8%/5.1% from GTA5 to

Cityscapes/BDDS/Mapillary. According to the comparison between model only with GTR (Tab. VIII b) and model only with LTR (Tab. VIII c), we can see that the proposed method LTR performs better than GTR in all three settings (G→C, G→B, G→M). We consider to combine GTR and LTR in our approach aiming to lead both global and local texture styles to our network. Compared to (Tab. VIII b) and (Tab. VIII c) respectively, model adopted GTR and LTR (Tab. VIII d) achieves remarkable results and gives a further improvement in all experimental setting. It means our mainly proposed two designs (GTR and LTR) can complement with each other. Moreover, it is important to find out the key factors affecting the performance of LTR. Several experiments are conducted to explore more combinations of painting styles in LTR. As shown

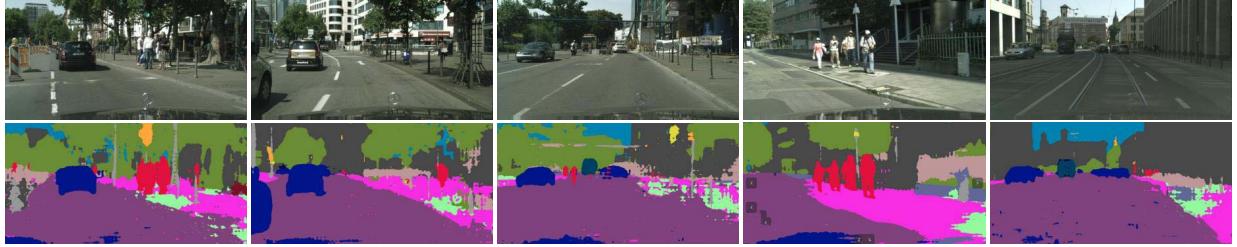


Fig. 14. Some failure examples of our framework. There is a misclassification when the sidewalk shows a similar appearance with terrain. (The figure is best viewed in color.)

TABLE VI

EXPERIMENT RESULTS FROM GTA5 TO CITYSCAPES. THE BEST RESULTS ARE MARKED IN BOLD. DA AND DG DENOTE DOMAIN ADAPTION AND DOMAIN GENERALIZATION RESPECTIVELY

| Network | DA / DG | Method | Access Tgt | mIoU |
|------------|---------|------------------|------------|-------------|
| VGG-16 | DA | FCN wild [8] | ✓ | 27.1 |
| | | CDA [9] | ✓ | 28.9 |
| | | CycADA [7] | ✓ | 34.8 |
| | | ROAD [32] | ✓ | 35.9 |
| | | I2I [43] | ✓ | 31.8 |
| | | AdaptSegNet [37] | ✓ | 35.0 |
| | | SSF-DAN [46] | ✓ | 37.7 |
| | | DCAN [45] | ✓ | 36.2 |
| | | CBST [44] | ✓ | 30.9 |
| | | CLAN [39] | ✓ | 36.6 |
| | DG | ADVENT [40] | ✓ | 36.1 |
| | DG | DPR [41] | ✓ | 37.5 |
| | DG | Ours | ✗ | 37.2 |
| ResNet-101 | DA | CycADA [7] | ✓ | 42.7 |
| | | ROAD [32] | ✓ | 39.4 |
| | | I2I [43] | ✓ | 35.4 |
| | | AdaptSegNet [37] | ✓ | 41.4 |
| | | DCAN [45] | ✓ | 41.7 |
| | | CLAN [39] | ✓ | 43.2 |
| | DG | ADVENT [40] | ✓ | 43.8 |
| | DG | DPR [41] | ✓ | 46.5 |
| | DG | Ours | ✗ | 43.7 |

TABLE VII

EXPERIMENT RESULTS FROM SYNTHIA TO CITYSCAPES. THE BEST RESULTS ARE MARKED IN BOLD. DA AND DG DENOTE DOMAIN ADAPTION AND DOMAIN GENERALIZATION

| Network | DA / DG | Method | Access Tgt | mIoU |
|---------|---------|--------------|------------|-------------|
| VGG-16 | DA | FCN wild [8] | ✓ | 20.2 |
| | | CDA [9] | ✓ | 29.0 |
| | | ROAD [32] | ✓ | 36.2 |
| | | DCAN [45] | ✓ | 35.4 |
| | | CBST [44] | ✓ | 35.4 |
| | | ADVENT [40] | ✓ | 31.4 |
| | | DPR [41] | ✓ | 33.7 |
| | DG | Ours | ✗ | 35.8 |
| | DA | ADVENT [40] | ✓ | 40.8 |
| | DG | DPR [41] | ✓ | 40.0 |
| | DG | Ours | ✗ | 39.7 |

in Tab. IX, adding painting styles is unable to boost the performance but obviously degrade it. We observe that adding painting styles also decreases the proportion of the raw style in image. Tab. X shows the ablation results of LTR with different component proportions. It is worthy noting that we utilize the single style image in LTR. Compared to LTR with multiple painting styles (Tab. IX b/c), LTR with single painting style (Tab. X d/e) achieves almost the same performances when the proportion of raw style is same. It illustrates that the proportion

TABLE VIII

PERFORMANCE CONTRIBUTION OF EACH PROPOSED DESIGN. THIS EVALUATION IS CONDUCTED FROM GTA5 (G) TO CITYSCAPES (C), BDDS (B) AND MAPILLARY (M). THE BEST RESULTS ARE IN BOLD

| Methods | TCPS | GTR | LTR | CGL | mIoU | | |
|---------|-------------------|-----|-----|-----|------|-------------|-------------|
| | | | | | G→C | G→B | G→M |
| (a) | ResNet-101 | | | | 34.0 | 28.1 | 28.6 |
| (b) | +GTR | | ✓ | | 39.2 | 35.4 | 32.1 |
| (c) | +LTR | | | ✓ | 41.8 | 36.1 | 33.7 |
| (d) | +GTR+LTR | | ✓ | ✓ | 42.4 | 38.6 | 37.8 |
| (e) | +GTR+LTR+CGL | | ✓ | ✓ | 43.2 | 39.1 | 38.4 |
| (f) | +TCPS+GTR | ✓ | ✓ | | 40.0 | 36.2 | 33.4 |
| (g) | +TCPS+LTR | ✓ | | ✓ | 42.4 | 36.8 | 34.8 |
| (h) | +TCPS+GTR+LTR | ✓ | ✓ | ✓ | 43.0 | 39.1 | 38.5 |
| (i) | +TCPS+GTR+LTR+CGL | ✓ | ✓ | ✓ | 43.7 | 39.6 | 39.1 |

TABLE IX

THE COMPARISONS OF LTR WITH DIFFERENT STYLE COMBINATIONS. THE BEST RESULTS OF EACH SRSS SETTING ARE MARKED IN BOLD

| Style Combination | mIoU | | | |
|---------------------------------------|-------------|-------------|-------------|-------------|
| | G→C | G→B | G→M | Avg |
| (a) raw style + one painting style | 43.7 | 39.6 | 39.1 | 40.8 |
| (b) raw style + two painting styles | 42.6 | 37.3 | 37.8 | 39.2 |
| (c) raw style + three painting styles | 41.9 | 37.0 | 37.3 | 38.7 |

TABLE X

THE COMPARISONS OF LTR WITH DIFFERENT COMPONENT PROPORTIONS. THE BEST RESULTS OF EACH SRSS SETTING ARE MARKED IN BOLD

| Proportion (raw style : single painting style) | mIoU | | | |
|---|-------------|-------------|-------------|-------------|
| | G→C | G→B | G→M | Avg |
| (a) 3 : 1 | 41.7 | 37.3 | 36.8 | 38.6 |
| (b) 2 : 1 | 42.4 | 37.8 | 38.1 | 39.4 |
| (c) 1 : 1 | 43.7 | 39.6 | 39.1 | 40.8 |
| (d) 1 : 2 | 42.7 | 37.3 | 37.6 | 39.2 |
| (e) 1 : 3 | 42.0 | 36.8 | 37.3 | 38.7 |

TABLE XI

THE COMPARISON BETWEEN CGL AND ITS DERIVATIVES. THE BEST RESULTS OF EACH SRSS SETTING ARE MARKED IN BOLD

| Consistency Strategy | mIoU | | | |
|----------------------|-------------|-------------|-------------|-------------|
| | G→C | G→B | G→M | Avg |
| CGL | 43.7 | 39.6 | 39.1 | 40.8 |
| CGS | 43.4 | 39.2 | 38.8 | 40.5 |
| CLS | 43.6 | 39.6 | 38.4 | 40.5 |
| CTF | 43.8 | 39.2 | 38.9 | 40.6 |

of the raw style is the main factor to affect the performance of LTR. As shown in Tab. IX, our approach achieves the best performance when the raw style has the same proportion with painting styles.

TABLE XII

PERFORMANCE COMPARISON BETWEEN UNREAL PAINTINGS AND REAL (OR SYNTHETIC) IMAGES. THE BEST RESULTS OF EACH SRSS SETTING ARE MARKED IN BOLD

| SRSS Setting | Type of Style Image | mIoU |
|--------------------|------------------------|-------------|
| GTA5 to Cityscapes | Paintings | 43.7 |
| | Images from Mapillary | 39.3 |
| | Images from BDDS | 40.7 |
| | Images from SYNTHIA | 36.8 |
| GTA5 to BDDS | Paintings | 39.6 |
| | Images from Cityscapes | 37.5 |
| | Images from Mapillary | 36.7 |
| | Images from SYNTHIA | 33.6 |
| GTA5 to Mapillary | Paintings | 39.1 |
| | Images from BDDS | 35.5 |
| | Images from Cityscapes | 35.8 |
| | Images from SYNTHIA | 32.9 |

3) Effects of Consistency Between GTR and LTR (CGL):

When further adding CGL to our framework, the performance of our framework (Tab. VIII e) is booted to 43.2%/39.1%/38.4% on G→C/G→B/G→M. Compared to the model with GTR and LTR (Tab. VIII d), although the CGL has little improvements, it has a widely stable performance gains in all settings. Besides, we show results for different consistent strategies in Tab. XI. The experiments include: Consistency between GTR and Source flow (CGS), Consistency between LTR and Source flow (CLS) and Consistency among Three Flows (CTF). As shown in Tab. XI, CGL shows the best performance on most of evaluation metrics.

4) Effects of Texture Complexity Based Painting Selection (TCPS):

Ablations with and without TCPS are performed to verify the proposed painting selection method. Note that when not adopting TCPS, 15 paintings are randomly selected from the painting dataset. Ablation results are given by taking the average value of 5 times experiments. The random painting selection of each experiment is independent. Compared with directly random selection from the painting dataset (Tab. VIII b-e), the model with TCPS (Tab. VIII h-i) improves the performance in each ablation setting by 0.5%~ 1.3%. This confirms the effectiveness of our TCPS.

5) Unreal Painting vs. Real or Synthetic Image:

In this paper, there are three types of datasets: Unreal Painting (“Painter by Numbers”), Synthetic Image (GTA5 and SYNTHIA) and Real Image (Cityscapes, BDDS and Mapillary). We mainly demonstrate the domain generalization ability of network can be significantly improved with unreal paintings. Whether it can meet the equal level of performance when using real or synthetic images to execute GTR and LTR? As shown in Tab. XII, we conduct experiments to address the above concern. It can be seen that the utilization of paintings outperforms using real and synthetic images in all SRSS settings. This is due to unreal paintings of dataset “Painting by Numbers” introduce various irregular and indistinct textures, enforcing the network to learn from other domain-invariant cues. While in real or synthetic dataset, the source images are transferred to a new domain. And the performance mainly depends on the similarity between the transferred domain and target domain.

TABLE XIII

PARAMETER ANALYSIS TOWARDS THE NUMBER OF TRAINING ITERATIONS. THIS EVALUATION IS CONDUCTED FROM GTA5 (G) TO CITYSCAPES (C), BDDS (B) AND MAPILLARY (M). THE BEST RESULTS OF EACH SRSS SETTING ARE MARKED IN BOLD

| Iterations | mIoU | | | |
|------------|-------------|-------------|-------------|-------------|
| | G→C | G→B | G→M | Avg |
| 100000 | 38.9 | 36.3 | 31.3 | 35.5 |
| 150000 | 42.0 | 37.8 | 38.3 | 39.4 |
| 200000 | 43.7 | 39.6 | 39.1 | 40.8 |
| 250000 | 43.4 | 39.1 | 38.8 | 40.4 |
| 300000 | 43.3 | 38.8 | 38.7 | 40.3 |

TABLE XIV

PARAMETER ANALYSIS OF λ IN EQUATION 7. THE REMARKABLE RESULTS ARE MARKED IN BLUE WHILE THE REST ARE MARKED IN GREEN

| λ | mIoU | | | |
|-----------|------|------|------|-------------|
| | G→C | G→B | G→M | Avg |
| 2 | 40.3 | 36.3 | 33.4 | 36.7 |
| 4 | 41.9 | 38.4 | 37.2 | 39.2 |
| 6 | 42.7 | 38.6 | 37.5 | 39.6 |
| 8 | 43.2 | 39.0 | 38.1 | 40.1 |
| 10 | 43.4 | 39.3 | 38.7 | 40.5 |
| 12 | 43.1 | 38.9 | 38.3 | 40.1 |
| 14 | 43.0 | 38.9 | 38.1 | 40.0 |
| 16 | 42.8 | 38.5 | 37.7 | 39.7 |
| 18 | 40.6 | 36.5 | 33.7 | 36.9 |
| 20 | 40.1 | 36.4 | 33.5 | 36.7 |

TABLE XV

PARAMETER ANALYSIS OF β IN EQUATION 13. THE BEST RESULTS OF EACH SRSS SETTING ARE MARKED IN BOLD

| β | mIoU | | | |
|---------|-------------|-------------|-------------|-------------|
| | G→C | G→B | G→M | Avg |
| 1e-3 | 42.8 | 39.3 | 36.9 | 39.6 |
| 1e-4 | 43.1 | 39.4 | 38.7 | 40.4 |
| 1e-5 | 43.7 | 39.6 | 39.1 | 40.8 |
| 1e-6 | 42.9 | 39.5 | 38.8 | 40.4 |
| 1e-7 | 42.5 | 39.1 | 37.4 | 39.7 |

6) Analysis of Parameters: The performances of our approach with different training iterations are shown in Tab. XIII. It can be seen that setting the number to 200000 produces the best results. Applying a larger or smaller iteration number leads to a degradation of the domain generalization performance. The λ in Equation 7 is the parameter to control the standard deviation of Gaussian kernel. We evaluate the performance of our method to this parameter. The results are recorded in Tab. XIV. For scores which are marked in blue, we can see that any two neighboring results are extremely close to each other (within 1%). In addition, all blue scores show outstanding performance and surpass the green scores in a large margin. The scores marked in blue are results of λ varying from 4 to 16. In order to increase the diversity of paintings while ensuring the performance, we randomly choose λ from [4, 16] in each iteration. The β in Equation 13 is the parameter to trade off the quality of segmentation loss with consistency loss. As shown in Tab. XV, we can see that when $\beta = 1e - 5$, our method achieves the best performance. The threshold of texture complexity T_{exc} has

TABLE XVI
PARAMETER ANALYSIS TOWARDS TEXTURE COMPLEXITY T_{exc} OF
SELECTED PAINTINGS

| T_{exc} | mIoU | | | |
|-----------|------|------|------|------|
| | G→C | G→B | G→M | Avg |
| 0.45 | 41.6 | 37.3 | 36.7 | 38.5 |
| 0.5 | 42.4 | 38.0 | 36.9 | 39.1 |
| 0.55 | 43.0 | 39.0 | 38.2 | 40.1 |
| 0.6 | 43.4 | 39.3 | 38.8 | 40.5 |
| 0.65 | 43.2 | 39.2 | 38.5 | 40.3 |
| 0.7 | 41.8 | 38.2 | 36.6 | 38.9 |
| 0.75 | 41.0 | 36.8 | 35.7 | 37.8 |

the same analysis trends with λ . From Tab. XVI, one can observe that when $T_{exc} \in [0.55, 0.65]$, our method achieves outstanding performances.

7) *Quantitative Results:* Fig. 13 provides some visual examples of the semantic segmentation results under three Cross-Domain settings. We can see that the proposed framework (the 5-th column) can generates segmentation with more details (e.g., pole, traffic sign and traffic light). Comparing results in the 2-nd column and the 3-rd column, the model with GTR produces less noisy regions, especially in the category of road. In addition, the usage of LTR brings a better boundary decision in various settings. For example, segmentation boundaries of car, pole, bicycle, traffic light and traffic sigh in the 4-th column are clearer than the first few columns. Fig. 14 shows some failure examples. One can observe that when the sidewalk has a dark green texture appearance, our method may misclassify some regions of sidewalk into terrain. The example in the 2-nd column shows that the shadow in sidewalk are misclassified, but human can easily recognize the shadow. This is due to that our model can not completely eliminate the effect of texture even using GTR and LTR.

VI. CONCLUSION

In this paper, we mainly present two texture randomization mechanisms for Domain Generalization. To relieve the network's strong bias towards recognizing texture, we propose a mechanism named Global Texture Randomization (GTR) to learn domain-invariant information. Next, Local Texture Randomization (LTR) mechanism is proposed to address the cases where only local regions are texture-different between the synthetic and real-world images. Moreover, to harmonize the two randomization mechanisms, consistency between GTR and LTR (CGL) is introduced. Besides, Texture Complexity based Painting Selection (TCPS) is proposed to ensure paintings are reliable enough for the above texture randomization mechanisms. Extensive experiments indicate that our approach achieves the state-of-the-art results in multiple synthetic-to-real dataset settings with different network backbones, which clearly demonstrates the effectiveness of our method.

REFERENCES

- [1] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, “Cascaded context pyramid for full-resolution 3D semantic scene completion,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7801–7810.
- [2] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
- [4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.
- [5] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, “SceneNet: Understanding real world indoor scenes with synthetic data,” 2015, *arXiv:1511.07041*. [Online]. Available: <http://arxiv.org/abs/1511.07041>
- [6] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, “Playing for data: Ground truth from computer games,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 102–118.
- [7] J. Hoffman *et al.*, “CyCADA: Cycle-consistent adversarial domain adaptation,” 2017, *arXiv:1711.03213*. [Online]. Available: <http://arxiv.org/abs/1711.03213>
- [8] J. Hoffman, D. Wang, F. Yu, and T. Darrell, “FCNs in the wild: Pixel-level adversarial and constraint-based adaptation,” 2016, *arXiv:1612.02649*. [Online]. Available: <http://arxiv.org/abs/1612.02649>
- [9] Y. Zhang, P. David, and B. Gong, “Curriculum domain adaptation for semantic segmentation of urban scenes,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2020–2030.
- [10] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, “Learning to generalize: Meta-learning for domain generalization,” in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3490–3497.
- [11] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, “MetaReg: Towards domain generalization using meta-regularization,” in *Proc. Adv. Neural Inf. Process. Syst. (NuerIPS)*, 2018, pp. 998–1008.
- [12] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, “Domain generalization for object recognition with multi-task autoencoders,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2551–2559.
- [13] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 23–30.
- [14] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” 2018, *arXiv:1811.12231*. [Online]. Available: <http://arxiv.org/abs/1811.12231>
- [15] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3722–3731.
- [16] V. Kolmogorov and R. Zabin, “What energy functions can be minimized via graph cuts?” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.
- [17] T. Meltzer, C. Yanover, and Y. Weiss, “Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation,” in *Proc. Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 428–435.
- [18] V. Kolmogorov, “A new look at reweighted message passing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 919–930, May 2015.
- [19] L. Grady and G. Funka-Lea, “Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials,” in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCV Workshops)*. Cham, Switzerland: Springer, 2004, pp. 230–245.
- [20] A. K. Sinop and L. Grady, “A seeded image segmentation framework unifying graph cuts and random Walker which yields a new algorithm,” in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
- [21] X. P. Dong, J. B. Shen, L. Shao, and L. van Gool, “Sub-Markov random walk for image segmentation,” *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 516–527, Feb. 2016.
- [22] A. Blake, P. Kohli, and C. Rother, *Markov Random Fields for Vision and Image Processing*. Cambridge, MA, USA: MIT Press, 2011.
- [23] J. Shen, X. Dong, X. Jin, L. Shao, F. Porikli, and J. Peng, “Submodular function optimization for motion clustering and image segmentation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2637–2649, Sep. 2019.
- [24] J. Shen, J. Peng, X. Dong, L. Shao, and F. Porikli, “Higher order energies for image segmentation,” *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4911–4922, Oct. 2017.

- [25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: <http://arxiv.org/abs/1511.07122>
- [26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: <http://arxiv.org/abs/1412.7062>
- [27] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1377–1385.
- [28] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.
- [29] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.
- [30] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4230–4239.
- [31] Y. Huang, Q. Wang, W. Jia, and X. He, "See more than once—Kernel-sharing atrous convolution for semantic segmentation," 2019, *arXiv:1908.09443*. [Online]. Available: <http://arxiv.org/abs/1908.09443>
- [32] Y. Chen, W. Li, and L. V. Gool, "ROAD: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7892–7901.
- [33] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NuerIPS)*, 2016, pp. 136–144.
- [34] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014, *arXiv:1409.7495*. [Online]. Available: <http://arxiv.org/abs/1409.7495>
- [35] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulo, "Auto-DIAL: Automatic domain alignment layers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5077–5085.
- [36] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C.-F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1992–2001.
- [37] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.
- [38] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.
- [39] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.
- [40] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.
- [41] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.
- [42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [43] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.
- [44] Y. Zou, Z. Yu, B. V. K. Vijaya Kumar, and J. Wang, "Domain adaptation for semantic segmentation via class-balanced self-training," 2018, *arXiv:1810.07911*. [Online]. Available: <http://arxiv.org/abs/1810.07911>
- [45] Z. Wu *et al.*, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 518–534.
- [46] L. Du *et al.*, "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 982–991.
- [47] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 10–18.
- [48] C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 87–97.
- [49] B. Gong, K. Grauman, and F. Sha, "Reshaping visual datasets for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst. (NuerIPS)*, 2013, pp. 1286–1294.
- [50] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.
- [51] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5542–5550.
- [52] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2100–2110.
- [53] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 464–479.
- [54] X. Pan, X. Zhan, J. Shi, X. Tang, and P. Luo, "Switchable whitening for deep representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1863–1871.
- [55] L. Huang, L. Huang, D. Yang, B. Lang, and J. Deng, "Decorrelated batch normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 791–800.
- [56] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst. (NuerIPS)*, 2017, pp. 386–396.
- [57] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: <http://arxiv.org/abs/1607.08022>
- [58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: <http://arxiv.org/abs/1502.03167>
- [59] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3642–3649.
- [60] I. Sato, H. Nishimura, and K. Yokoi, "APAC: Augmented PAttern classification with neural networks," 2015, *arXiv:1505.03229*. [Online]. Available: <http://arxiv.org/abs/1505.03229>
- [61] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1058–1066.
- [62] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2003, pp. 958–963.
- [63] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.
- [64] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid, "A Bayesian data augmentation approach for learning deep models," in *Proc. Adv. Neural Inf. Process. Syst. (NuerIPS)*, 2017, pp. 2797–2806.
- [65] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," 2017, *arXiv:1702.05538*. [Online]. Available: <http://arxiv.org/abs/1702.05538>
- [66] T. Dreossi, S. Ghosh, X. Yue, K. Keutzer, A. Sangiovanni-Vincentelli, and S. A. Seshia, "Counterexample-guided data augmentation," 2018, *arXiv:1805.06962*. [Online]. Available: <http://arxiv.org/abs/1805.06962>
- [67] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*. [Online]. Available: <http://arxiv.org/abs/1710.09412>
- [68] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*. [Online]. Available: <http://arxiv.org/abs/1708.04552>
- [69] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.
- [70] U. Köthe, "Edge and junction detection with an improved structure tensor," in *Proc. Joint Pattern Recognit. Symp. (JPRS)*. Cham, Switzerland: Springer, 2003, pp. 25–32.
- [71] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

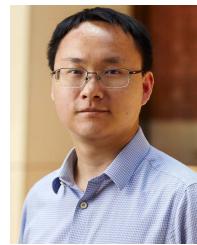
- [72] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," 2019, *arXiv:1906.01916*. [Online]. Available: <http://arxiv.org/abs/1906.01916>
- [73] L. C. Andrews, *Special Functions of Mathematics for Engineers*, vol. 49. Bellingham, WA, USA: SPIE, 1998.
- [74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [76] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. Adv. Neural Inf. Process. Syst. Workshops (NeurIPS Workshops)*, Long Beach, CA, USA, 2017.
- [77] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.
- [78] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
- [79] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.
- [80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [81] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105.
- [82] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: <http://arxiv.org/abs/1706.05587>



Duo Peng received the B.Eng. degree in electronic information engineering from Sichuan University (SCU), Chengdu, China, in 2019, where he is currently pursuing the M.S. degree in information and communication engineering. His main research interests include deep learning, transfer learning, and semantic segmentation.



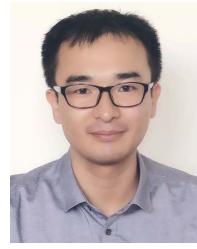
Yinjie Lei (Member, IEEE) received the M.S. degree in image processing from Sichuan University (SCU), China, in 2009, and the Ph.D. degree in computer vision from The University of Western Australia (UWA), Australia, in 2013. Since 2017, he has been serving as the Vice Dean of the College of Electronics and Information Engineering, SCU, where he is currently an Associate Professor. His main research interests include 3D biometrics, object recognition, and semantic segmentation.



Lingqiao Liu is currently a Senior Lecturer with the University of Adelaide and the Australian Institute for Machine Learning. He received the B.S. and M.S. degrees in communication engineering from the University of Electronic Science and Technology of China, Chengdu, in 2006 and 2009, respectively, and the Ph.D. degree from the Australian National University, Canberra, in 2014. In 2016, he was awarded the Discovery Early Career Researcher Award from the Australian Research Council and the University Research Fellow from the University of Adelaide. His current research interests include low-supervision learning and various topics in computer vision and natural language processing.



Pingping Zhang received the B.E. degree in mathematics and applied mathematics from Henan Normal University (HNU), Xinxiang, China, in 2012, and the Ph.D. degree in signal and information processing from the Dalian University of Technology (DUT), Dalian, China, in 2020. He is currently an Associate Professor with the School of Artificial Intelligence, DUT. He has authored over 30 journals/conference papers. His research interests include deep learning, saliency detection, object tracking, and semantic segmentation.



Jun Liu (Member, IEEE) received the B.Eng. degree from Central South University, the M.Sc. degree from Fudan University, and the Ph.D. degree from Nanyang Technological University. He is currently an Assistant Professor with the Singapore University of Technology and Design. He is an organizer of several international workshops. His works have been published in premier computer vision journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), CVPR, ICCV, and ECCV. His research interests include computer vision and artificial intelligence. He received the Best Thesis Award from EEE, Nanyang Technological University, and several best paper awards from PREMIA for the works on video analytics and human activity understanding. He is the Area Chair of ICLR 2022 and WACV 2022. He is an Associate Editor of *Neurocomputing* (Elsevier), *Signal, Image and Video Processing* (Springer), *IET Computer Vision*, and *Journal of Electronic Imaging* (SPIE), and a Guest Editor of *The Visual Computer Journal* (Springer).