

Bridging domain gaps in CNNs: a comprehensive approach with adaptation and randomization strategies

Chen Sun^{1,*}, Xingxin Chen¹, Yukun Lu¹, Dongpu Cao² and Amir Khajepour¹

¹Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, Canada

²School of Vehicle and Mobility, Tsinghua University, Beijing, China

* Correspondence author; E-mail: c87sun@uwaterloo.ca.

Abstract: Metric-based methods are promising approaches to domain adaptation, aiming to align the marginal distributions of different domains with similar conditional distributions. In traditional approaches, the metric function is manually designed to measure the distance across domains. Adversarial methods can be considered an automatic learning approach to the metric function. Instead of relying on the quality of the metric function, we outline a generalized framework for domain randomization, which first introduces moderate perturbations as a form of randomness and then combines the advantages of metric-based domain adaptation with domain randomization. We propose a novel approach leveraging simulation environments to generate extensive, annotated data for diverse scenarios. Our focus is on simulation-to-real transfer for semantic segmentation tasks, acknowledging CNN's texture bias. We introduce a domain randomization technique that limits meaningful texture information and devise a mapping function for image transformation. The proposed approach emphasizes geometry consistency and style transfer, providing a practical solution for efficient simulation to real-world transfer. The experiments are conducted on the domain generalization task from GTA to Cityscapes and BDD, and evaluated on the semantic segmentation performance. Our method achieves superior results in most segmentation classes compared with the benchmark models by 5.2 to 10 in terms of mIoU. Remarkably, our generalization strategy does not require access the target domain data at training time.

Keywords: machine learning; autonomous vehicles; domain transfer; robust learning

1. Introduction

Deep neural networks excel in tasks assuming data are independent and identically distributed (i.i.d.). Yet, the concept of *domain shift* [1] challenges this assumption, especially in varied cyber-physical-system applications like autonomous driving [2]. Such environmental changes and domain shifts critically hinder the real-world application of these systems [3].

Transfer learning typically involves using pre-trained models to extract features from large, annotated datasets, then fine-tuning on specific tasks [4, 5]. However, gathering and annotating enough data for fine-tuning is costly and time-intensive. Additionally, the vast amount of labeled data needed to encompass the diversity of real-time scenarios in applications such as autonomous driving and robotic control makes data collection not only prohibitively expensive but also potentially hazardous.

Simulation environments address many data collection issues by providing unlimited, well-annotated data, eliminating safety and privacy concerns, including those related to synthetic humans [6]. These environments can balance data distributions to reduce bias and offer interactive rewards for reinforcement learning. However, despite their advantages,



Copyright©2024 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited

simulations suffer from a reality gap with the real world. While game engine-based simulators like CARLA [7] and AirSim [8] require moderate computational resources, they fall short of delivering photo-realistic results.

A key strategy for bridging the simulation-reality gap is domain adaptation, which aligns data distributions across domains into a shared latent space, using metric-based regularization by the task model. Current methods primarily aim to minimize the distance between source and target distributions [5]. Early efforts focused on reducing the marginal and conditional distribution differences [9, 10]. Later, deep learning approaches used CNNs to map features to a common space [11, 12], and GAN-inspired techniques employed CNN-based domain discriminators for measuring domain distances [4, 13]. Additionally, to enhance training regularization, semantic transformations and batch spectral penalization were applied to ensure content consistency during domain mapping [14, 15].

Beyond minimizing differences at the pixel and feature levels, another line of thought considers instance-level differences [16]. The authors of [16] focus on minimizing the distribution differences in scene composition between simulation and reality, rather than aligning semantic and pixel distributions. Additional instance-level relationship information [17] and learning methodologies [18] also contribute to the transfer process. However, the additional graphic assets are not free, and the domain gap between simulation and reality is not solved merely by combining game engine assets with extracted scene composition distribution.

We classify these methods as metric-based, with a typical workflow shown in Figure 1. Metric-based domain adaptation depends heavily on the precision of metrics measuring inter-domain distances. Despite the high accuracy of CNN-based discriminators, these metrics are often insufficient. In contrast, existing domain randomization techniques, which rely on manual augmentations informed by prior domain knowledge, may not effectively identify a common feature space [19]. Recent advancements utilize style discriminators in style transfer-based domain randomization to measure domain distances [19]. Building on this, we propose a metric-guided domain randomization approach that integrates metric-based domain adaptation for enhanced accuracy and robustness in domain transfers.

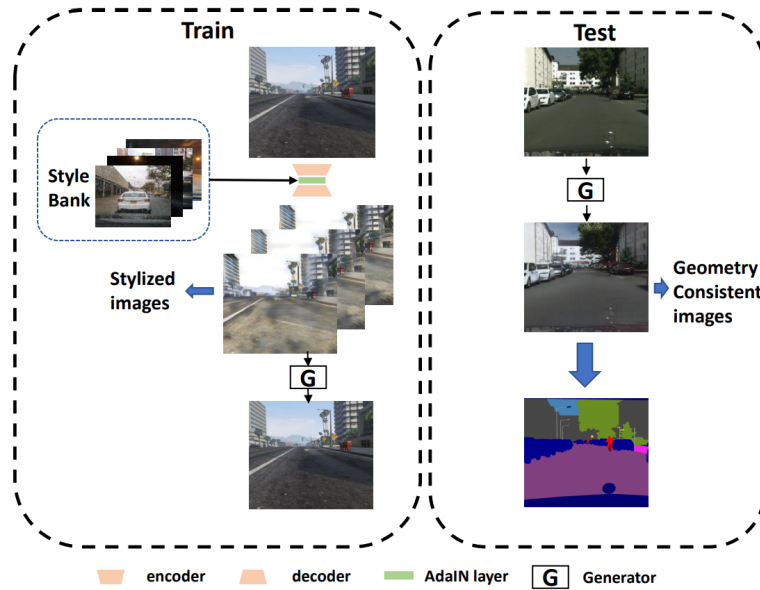


Figure 1. The typical pipeline of the metric-based domain adaptation.

A concise framework comparison between our proposed domain randomization scheme and traditional metric-based adaptation approaches is presented in Figure 2. Unlike metric-based domain adaptation methods, which heavily rely on metric design, our approach

initially combines multiple simple domain adaptation modules. These modules augment the source data’s marginal distribution toward the target data’s marginal distribution with selected perturbations. Additionally, we can track the level of perturbation and introduce a degree of randomness into the distribution transfer learning process. To ensure that only consistent information influences the training outcome, we employ a contrastive learning module. Our paper’s main contributions are as follows:

- We propose a novel framework for domain randomization, synergizing multiple domain adaptation networks to enhance task performance in the target domain effectively.
- We develop a domain generalization model with a triplet network, based on this framework, for sim-to-real adaptation. This model utilizes style transfer for domain randomization and contrastive learning to regulate the learning process.
- Our approach is straightforward to implement and generalizes across various real-world application domains. We demonstrate the efficacy of our model in the GTA5 to CityScapes and BDD100K semantic segmentation policy transfer tasks, showing comparable performance to other metric-based methods with enhanced generalization capabilities.

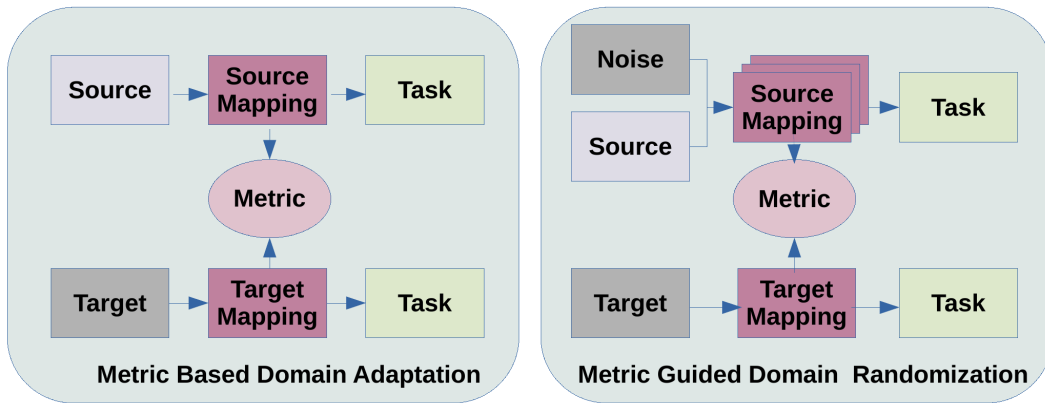


Figure 2. A brief framework comparison between popular metric-based domain adaptation (left) and the proposed metric-guided domain randomization (right).

2. Related works

2.1. Simulation-to-reality domain adaptation

Recent studies distinguish the simulation-reality gap into instance-level and pixel-level differences. The former relates to scenario complexity and distribution, while the latter concerns image quality and style across domains. Although much research, such as [14, 20–22], has focused on pixel-level gaps, recent efforts [16, 23] have begun addressing the instance-level gap, attracting significant attention.

Central to optimizing domain adaptation is the choice of distance metrics. Popular metrics include A-Distance [24] and Maximum Mean Discrepancy (MMD) [11, 12, 25], with MMD used to minimize distribution distances within CNNs and A-Distance challenging to compute during real-time training. Recent adversarial methods [13] approximate these metrics effectively, enhancing domain differentiation. Additionally, advances in domain adaptation metrics, like the cycle-consistent loss used in [14] to preserve semantic information, and Batch Spectral Penalization (BSP) loss [15], which balances transferability and discriminability, underscore the importance of targeted regularization and perturbation levels in domain adaptation.

2.2. Source-free domain adaptation

The other, more challenging setting for domain adaptation considers the absence of source data, often due to privacy or storage limitations, which traditionally plays a crucial role in guiding the adaptation process. Recent works have been designed to overcome the challenges of adapting without source data, utilizing methods such as pseudo-labeling [26], adversarial training [27], and distribution alignment [28]. Augmenting the data with models or other expert knowledge and maximizing the use of the data represent the mainstream of research in this challenging problem setting [29]. Source-free domain adaptation continues to grow as a vital area of research, with broad applications in resource-constrained environments [30]. The authors encourage readers who are interested in this topic to refer to the recent survey [30]. Anchored by the main idea of pushing the boundaries of what is possible in domain adaptation, this paper considers randomization techniques, assuming the case where target data is limited.

2.3. Domain randomization

This strategy seeks domain invariant features through randomization rather than merely minimizing inter-domain distances. Traditional augmentation methods like flipping and rotation are commonly employed in domain randomization to enhance model generalization. [31] explores learning augmentation policies for sim-to-real transfers. Recent automatic augmentation methods [32, 33] are also applicable to domain randomization. Authors in [19] investigate style transfer [34] for consistent model styling. Unlike standard training with randomized data, Randomized-to-Canonical Adaptation Networks (RCANs) was used in [35], bridging the reality gap by translating randomized images into their original canonical forms.

3. Methods

Our approach can be divided into two parts, illustrated in Figure 3. The first part is domain randomization with neural style transfer: mapping the source domain images to different styles, such that, the target domain is just another randomized style. The second part is a randomized images to original images mapping structure utilizes a customized conditional generative adversarial network structure [36]. The relation between simulation, randomized simulation and reality is shown in Figure 4.

3.1. Metric guided domain randomization

We present the general framework for metric-guided domain randomization here. Consider we have access to source images X_s and labels Y_s drawn from a source domain distribution $p_s(x, y)$. In the sense of unsupervised adaptation, we have access to limited unlabelled target images X_t from target domain distribution $p_t(x, y)$ where there is no label but we aim to learn a task mapping function M to transfer knowledge. In [13], authors assume they have direct access to the target images. However, the limited target images often lead to model bias to the target domain distribution $p_t(x, y)$. Here we consider a different setting by adding random perturbations to approximate the target distribution. To consider the randomness of the perturbations, we denote the noise set as N and specific noise $n \in N$. The major goal of our proposed framework is to randomize and augment the source mapping to make sure $M(X_t, n)$ can not be distinguished from the demonstrated task in the source domain with a collection of randomness $\{M(X_s, n_i)\}_{i=1}^k$, where k is the size of the selected noise set. In this case, the source classifier C_s can be directly applied to the target domain, e.g., we have a unified classifier $C = C_s = C_t$.

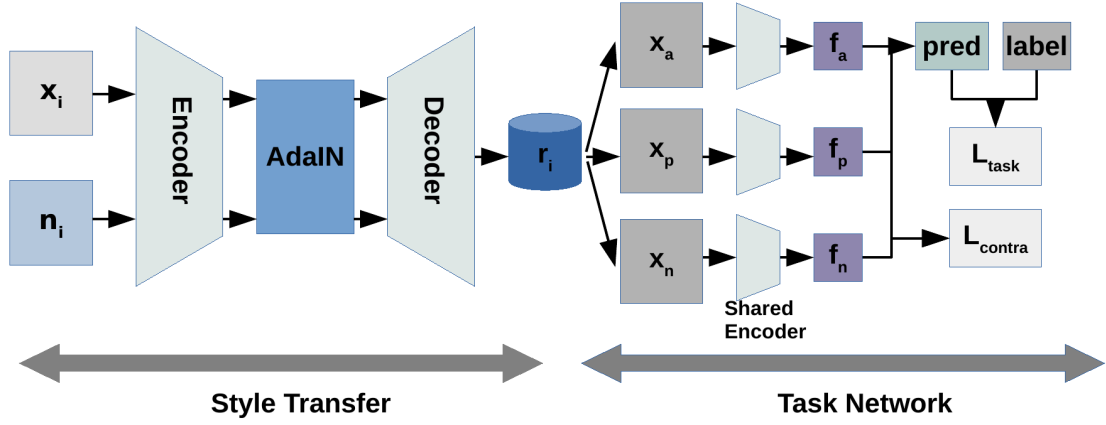


Figure 3. Overview of the Proposed Metric-Guided Domain Randomization Pipeline. This figure illustrates our pipeline, which begins with a style transfer network generating random stylized images from noise $n_i \in N$. These images are classified as positive x_p or negative x_n by comparing them with the anchor image x_a . The contrastive loss is employed in task policy training to ensure that feature mappings of positive images align closely with those of the anchor, while those of negative images differ significantly.

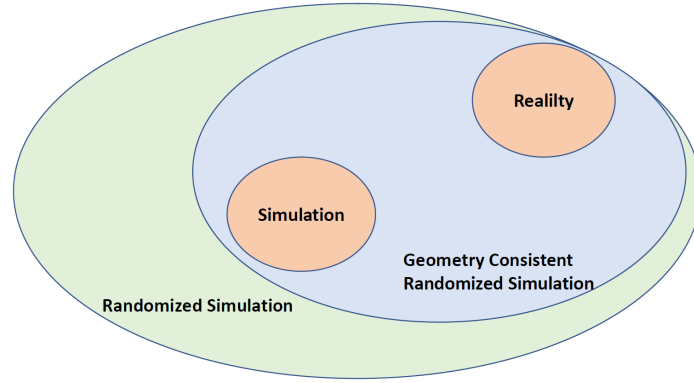


Figure 4. Geometry consistent domain randomization and common domain randomization.

The supervised loss for the source classification model can be defined as as

$$L_{\text{cls}}(X_s, Y_s, N) = \mathbb{E}_{(x_s, y_s, n) \sim (X_s, Y_s, N)} - \sum_{k=1}^K \mathbb{1}_{[k=y_s]} \log C(M(x_s, n)) \quad (1)$$

To make sure the selected perturbation noise N can effectively randomize the source domain mapping $M(X_s, N)$ instead of being ignored, we present noise classifier C_n with the supervised loss $L_n(N, M)$ as

$$L_n(X_s, N, M) = \mathbb{E}_{(x_s, n) \sim (X_s, N)} - \sum_{k=1}^K \mathbb{1}_{[k=n]} \log C_n(M(x_s, n)) \quad (2)$$

Finally, we need to make sure the task mapping in target domain with arbitrary perturbations $M(X_t, \tilde{n})$ can not be distinguished by a domain discriminator D , from the task model accepting the source domain input $M(X_s, N)$ for the transfer learning robustness. The supervised loss for discriminability L_D is defined as

$$L_D(X_s, X_t, M, N) = -\mathbb{E}_{(x_s, n) \sim (X_s, N)} [\log D(M(x_s, n))] - \mathbb{E}_{x_t \sim X_t} [\log(1 - D(M(x_t, \tilde{n})))] \quad (3)$$

where $\tilde{n}$ is a specific noise.

The generic optimization goal for the metric-guided domain randomization considering unknown perturbations is derived below:

$$\begin{aligned}
 & \min_{M,C} L_{\text{cls}}(X_s, Y_s, N) \\
 & \min_{M,C_n} L_n(X_s, N, M) \\
 & \min_M \max_D L_D(X_s, X_t, M, N)
 \end{aligned} \tag{4}$$

The proposed formulation of a metric-guided domain randomization problem can be considered as a simultaneous optimization problem which is often hard to solve for the global optimal solution. Especially, the direction of randomness N included will have an impact on the efficiency of domain transfer. Several problems remain for finding the solution of the optimization problem (4).

- How to find “proper” noise to the feature mapping?
- Should the introduced perturbation be the features from the target domain data, data similar to target data or just random noise?
- In terms of the training process, should we divide the whole pipeline into metric based domain randomization and training on randomized data, or one-staged adversarial training progress?

We will discuss these issues and propose our solution and designed model in the following sections.

3.2. Randomization based domain generalization

Our framework for domain randomization enhances adaptive learning methods in fields like autonomous driving and robotic vision, where semantic segmentation is critical for environmental understanding. Although source images and labels are available from public datasets and simulations, accessing target domain data can bias the model towards a limited distribution, undermining generalization. In domain adaptation, where target data are scarce, this issue is pronounced. Instead of relying on direct target data access, our model uses a basic understanding of geometric consistency. Recognizing that both simulation and real-world settings share geometric attributes, despite differences in texture and lighting, we employ a geometry-consistent domain randomization approach initially to randomize source data. Additionally, we implement a triplet structure in our model to focus learning on geometry-consistent information, enhancing its effectiveness.

To solve the union optimization problem in (4), we propose a dual-phase pipeline that integrates style transfer with a triplet network, underpinned by contrastive learning. The first phase uses style transfer, as shown in Figure 1, to address (2) through effective randomization techniques [19]. We employ adaptive instance normalization [37] to inject noise into feature mappings, carefully regulating the perturbation to retain essential content from the source image. Additionally, we use a style loss $L_n(N, M)$, detailed in (2), to maintain distinct noise characteristics in the images.

The discriminability loss (3) is a typical min-max two player problem where a GAN model is typically accepted in [13]. Although GAN should have excellent performance in Nash equilibrium, but only in the case of convex function can gradient descent guarantee Nash equilibrium. When both sides of a game are represented by neural networks, it is possible for them to always adjust their strategies without actually reaching equilibrium. Another point is that the learning process of GAN often suffer from collapse problem where the generator starts to degenerate. Considering these problems, we propose a simple-to-implement supervised training pipeline to handle (1) and (3) simultaneously. The task network stage illustrated in Figure 1 take advantage of a triplet structure to filter the augmentation with “proper” perturbations. We argue that the discriminability can be measured the consistency of the

semantic segmentation results. That is, we can use geometric consistency as an approximation measure of the model discriminability in (3). Moreover, we use a triplet network to tell whether the source image with injected perturbation $x_s \oplus n_i$ belongs to the source domain distribution, e.g., classify whether $x_s \oplus n_i \sim p_s(x, y)$ is true. The proposed triplet structure is very useful in making the task model robust on the geometric consistent information.

3.3. Style transfer for domain randomization

To introduce random noise to the source images while ensuring geometric consistency with the original images, we encountered a challenge. Directly applying a pixel-wise constraint, such as the L_1 loss, during training contradicted Equation (2) and rendered the perturbations indistinguishable from the randomized images. Consequently, we adopted a perceptual metric as suggested by [38] to measure the discrepancy between the source and randomized images. This process involves encoding both the content image c and the style image s into a feature space using a fixed encoder f . An adaptive instance normalization layer is then utilized to align the mean and variance of the content feature maps with those of the style feature maps:

$$t = \text{AdaIN}(f(c), f(s)). \quad (5)$$

A decoder g is used to map the target feature map back to target image space where

$$T(c, s) = g(t). \quad (6)$$

The total training loss is:

$$\theta_f^* = L_s + \lambda_1 L_c, \quad (7)$$

where

$$L_s = \sum_{i=1}^L \|\mu(\phi_i(g(t))) - \mu(\phi_i(s))\|_2 + \sum_{i=1}^L \|\sigma(\phi_i(g(t))) - \sigma(\phi_i(s))\|_2, \quad (8)$$

and

$$L_c = \|f(g(t)) - t\|_2. \quad (9)$$

In (8), ϕ_i denotes activation layers of f . While the adaptive normalization operation align noise into intermediate features, L_s is used to make sure n is still noticeable after reconstruction, as is described in equation (2).

3.4. Contrastive learning through triplet network

Our task network needs to learn from $M(x_s, n)$ and control the effects of n . We consider it as a contrastive learning problem and try to learn from the differences between different generated images (as shown in Figure 5). In this section we explore the use of triplet network and siamese network [39]. We generate different images of the same source image with different style images and every two of the images can be considered as a positive pair. To generate negative pairs, several methods can be used, including rotation, coloring and take temporal-shifted images as negative pairs. We add a triplet ranking loss to make sure our intermediate feature representation focuses on geometry information instead of image texture. For an anchor image x_a , positive image x_p has the same geometry information with x_a and negative image x_n has similar texture information with the anchor image. The triplet ranking loss is:

$$L_{\text{contrastive}}(x_a, x_p, x_n) = \max(0, m + \|f(x_a) - f(x_p)\| - \|f(x_a) - f(x_n)\|) \quad (10)$$

Or we can just use a pairwise ranking loss:

$$L_{\text{contrastive}}(x_a, x_p, x_n) = \begin{cases} \|f(x_a) - f(x_p)\| & \text{if } (x_a, x_p) \text{ is positive} \\ \max(0, m - \|x_a - x_n\|) & \text{if } (x_a, x_n) \text{ is negative} \end{cases} \quad (11)$$

Another choice of training objective is:

$$L_{task}(x_a, x_p, x_n) = L_{seg}(x_a, y_a) + \lambda L_{contrastive} \quad (12)$$

where f is a feature mapping function of our task network.



Figure 5. Contrastive learning triplets with real and stylized images. The ‘Anchor’ image is processed using a perceptual similarity metric. The ‘Positive’ image shares the same content as the Anchor but is stylized differently. In contrast, the ‘Negative’ image, while stylistically similar to the Anchor, is based on a content-wise similar yet distinct image.

In our research, we incorporated a spatial pyramid pooling layer prior to calculating the distance between feature representations, as suggested in the study by [19]. A critical aspect of contrastive learning is the selection of negative pairs. Our experiments highlighted the challenge in consistently selecting hard negative examples to ensure stable gradient provision. While Noise Contrastive Estimation (NCE), as outlined by [40], effectively combines triplet loss and softmax for such scenarios, its requirement for storing numerous negative pairs led us to focus solely on positive pairs in our experiments. Consequently, the pairwise ranking loss in our study aligns with the triplet loss formulation. However, as indicated by Theorem 2 in [41], no single arrangement of batch samples and learning configurations guarantees convergence to an optimal point. To address this, we carefully selected challenging negative examples and employed categorical cross-entropy loss to distinguish the positive sample amid a range of unrelated noise samples.

4. Experiments and evaluation

In this section, we describe the experimental setup and present results of our metric-guided domain randomization method applied to semantic segmentation tasks. Additionally, we analyze various experimental setups and compare our method with other approaches. Notably, our policy transfer process does not require any target domain data or information related to the target domain. The only prerequisite is prior knowledge that the current simulation environment differs from the real-world environment in aspects like texture, lighting, and physics, yet shares similarities in geometric information. This stringent experimental setup ensures the robust generalization ability of our proposed method.

4.1. Experimental settings

In our experiments, we utilized the FCN8s-VGG16 model, as detailed by [42], for our task baseline policy. This involved assessing the semantic segmentation capabilities of the model, which was trained on a synthesized dataset and then tested across various real-world datasets.

We selected the GTA5 dataset [43] as our synthetic dataset, and the Cityscapes [44] and BDD100K [45] datasets as our real-world datasets. The GTA5 dataset, sourced from the Unreal Engine, offers vehicle-centric views with highly accurate, pixel-wise semantic segmentation labels extracted directly from the game. It comprises 24,966 images with a resolution of 1914×1052 pixels, and the class labels are compatible with those in the CityScapes dataset. CityScapes, on the other hand, provides a diverse urban environment perspective, collected from 50 different European cities. It includes 5000 finely annotated frames and 20,000 weakly annotated frames, offering a resolution of 2048×1024 pixels and encompassing 30 classes, though our study only utilizes 19 of these classes.

The BDD100K dataset, gathered via dashcam, presents a diverse collection of over 10,000 images with both pixel-level and rich instance-level annotations, at a resolution of 1280×720 pixels. Through this varied selection of datasets, our study aims to rigorously evaluate the adaptability and effectiveness of the FCN8s-VGG16 model in handling diverse, real-world semantic segmentation tasks, leveraging the strengths of both synthesized and real-world data.

4.2. Training and implementation details

To enhance our model's performance in semantic segmentation, we trained an AdaIN network using the GTA5 dataset's training set for content images and the Painter by Numbers dataset for style images. This approach augmented the GTA5 training set with images created by the AdaIN model using varied style images from the Painter by Numbers collection. For every content image from the GTA5 dataset, two stylized images were generated, forming positive pairs. These images underwent a filtering process using a perceptual loss with a predetermined threshold to maintain quality.

Our primary model for semantic segmentation was the FCN8s-VGG16. The network's input comprised pairs of these generated images, each derived from the same source image in the GTA5 dataset. To optimize our model, we trained all semantic segmentation models using images resized to a resolution of 512×1024 pixels. The training process involved the use of contrastive loss, which was computed across various activation layers of the FCN8s model. This methodology aimed to leverage the rich, diverse data from the GTA5 and Painter by Numbers datasets, synthesizing a robust training environment that would enhance the model's ability to generalize and perform effectively in semantic segmentation tasks.

For a more diverse comparison with prevalent methods, we selected VGG-16 as the foundation backbone for our FCN. These networks' feature extraction layers are initialized using pretrained ImageNet models. The network is built on PyTorch, employs the Adam optimizer with a batch size of 16 for both the baseline and our models.

4.3. Results evaluation

Our evaluation protocol followed the guidelines in [46]. Real-world RGB images were resized to 512×1024 before being fed into the model trained with synthesized images. The output semantic predictions were then resized back to their original resolution for evaluation using the standard Pascal VOC intersection-over-union (IoU) metric. We employed the mean IoU (mIoU) across all categories as our evaluation metric. The model trained on the augmented GTA5 dataset was tested using the CityScapes dataset, with ablation study results shown in Table 1.

4.3.1. Qualitative analysis and ablation study

We evaluated our proposed method's performance using GTA5 as the source of labeled training data, and conducted assessments on Cityscapes' validation set. Our analysis included comparisons between two foundational semantic segmentation frameworks: FCN-VGG16

(A) and FCN-VGG19 (B), with “C” in the target column in Table 1, denoting the Cityscapes dataset and “B” representing the BDD100K dataset. This approach allowed us to gauge our model’s effectiveness across different architectures and target datasets.

Table 1. Performance evaluation on GTA5 dataset.

Method		Generalize from GTA5																				
		target architecture	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
FCN wld [22]	C A	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1	
AdaSeg [47]	C A	72.9	30.0	74.9	12.1	13.2	15.3	16.8	14.1	79.3	14.5	75.5	35.7	10	62.1	20.6	19	0.0	19.3	12	31.4	
MCD [48]	C A	86.4	8.5	76.1	18.6	9.7	14.9	7.8	0.6	82.8	32.7	71.4	25.2	1.1	76.3	16.1	17.1	1.4	0.2	0.0	28.8	
FSDR [49]	C A	73.2	30.1	68.8	17.4	3.9	18.1	3.3	0.3	79.9	25.1	72.9	24.5	0.1	70.5	11.2	14.9	0.2	0.2	0.3	24.1	
TLDR [50]	C A	76.8	30.5	71.1	20.6	5.6	19.2	6.5	0.5	81.1	28.6	75.3	26.3	0.1	74.4	15.5	17.4	0.4	0.3	0.5	26.3	
Source	C A	8.3	10.3	37.2	4.0	12.9	14.6	8.4	3.4	74.6	9.9	31.3	36.5	0.1	53.4	9.5	6.1	0.6	0.1	0.0	16.9	
Proposed	C A	77.9	31.5	72.0	10.2	0.0	19.3	2.1	0.0	79.6	23.6	76.9	28.6	0.0	60.4	8.6	6.2	0.0	0.0	0.0	26.2	
Source	C B	7.4	10.7	35.0	3.8	10.7	13.4	13.8	2.4	76.0	14.2	30.2	40.6	0.0	48.3	9.3	4.2	2.0	0.8	0.0	17.0	
Proposed	C B	82.6	32.4	74.2	16.4	0.0	22.9	0.2	0.0	81.1	21.2	77.1	33.6	0.0	61.4	10.4	2.7	0.0	0.2	0.0	27.2	
Source	B A	21.3	15.4	42.7	1.8	9.9	19.0	13.2	2.7	52.4	11.9	74.3	27.0	0.0	45.9	5.9	7.0	0.0	0.3	0.0	18.4	
Proposed	B A	54.8	21.1	60.0	4.3	0.0	24.5	8.4	0.0	68.1	21.7	84.4	28.6	0.0	53.0	8.9	9.6	0.0	0.0	0.0	23.6	

Overall, the proposed method helps to boost the overall mIoU performance for GTA-Cityscapes by 10, and around 5.2 for the BDD100K dataset. The impact of resolution on domain adaptation performance is discussed in previous studies [51]. Generally, for metric-based domain adaptation methods, higher resolution leads to better performance in both source and target domains. In our experiments, however, while high resolution improves performance in the randomized source domain, it does not significantly enhance target domain performance. For instance, images from the GTA5 dataset resized to 512×1024 and then randomly cropped to 512×512 yield a target domain mIoU (FCN8s-VGG16) of 27.4, compared to 27.1 for the larger resolution. The performance in the source domain semantic segmentation decreases with lower resolution, dropping from 34.7 to 40.1. Moreover, as noted in [51], even with the same base model, reported performance in the source domain varies across studies due to implementation differences. Changing the base model from VGG16 to VGG19 in our method does not improve performance, suggesting that the bottleneck lies in the style transfer module. Figures 6 and 7 showcased some examples of the semantic segmentation results on the semantic segmentation task adapting from GTA to CityScape and BDD100K respectively.

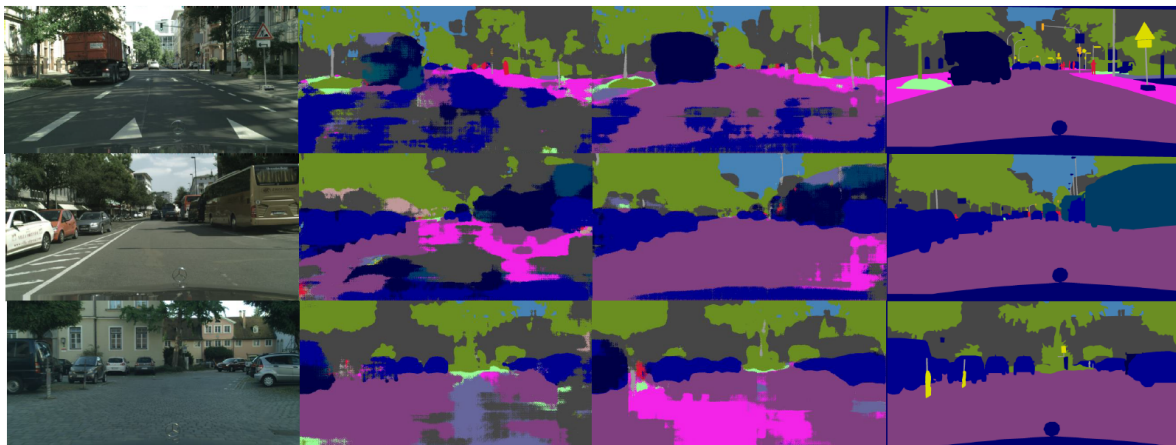


Figure 6. Qualitative semantic segmentation results of the generalization from GTA to CityScapes. The first column is the original RGB image, the second column is the result of FCN8s-VGG16 model trained directly on GTA5 dataset (baseline), the third column is the result of our proposed method and the last column is the ground truth.

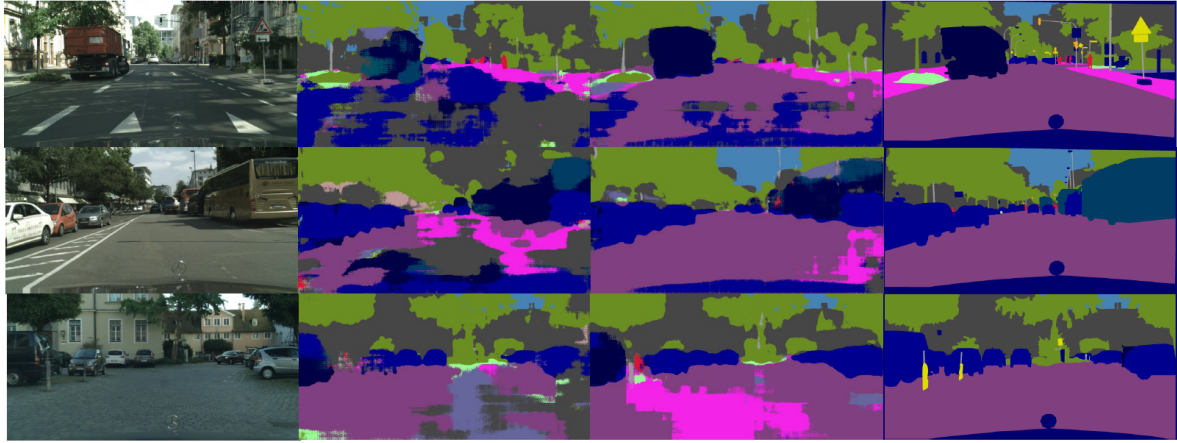


Figure 7. Qualitative semantic segmentation results of the generalization from GTA to BDD100K. The first column is the original RGB image, the second column is the result of FCN8s-VGG16 model trained directly on GTA5 dataset (baseline), the third column is the result of our proposed method and the last column is the ground truth.

4.3.2. Geometric consistency limitations

In our experiments, we observed that the style image becomes nearly indistinguishable from the randomized image, as demonstrated in Figure 8. While this result might be problematic for typical style transfer tasks, it aligns with our requirements. The geometric consistency property exists in many real-world applications that need domain adaptation. For example, applications such as industrial automation deployed in the controlled environment [52] or those objects with well-defined geometric properties such as driving vehicles and pedestrians [53]. Studies in domain adaptation and simulation often highlight scenarios where geometric consistency aligns closely with real-world conditions [54].

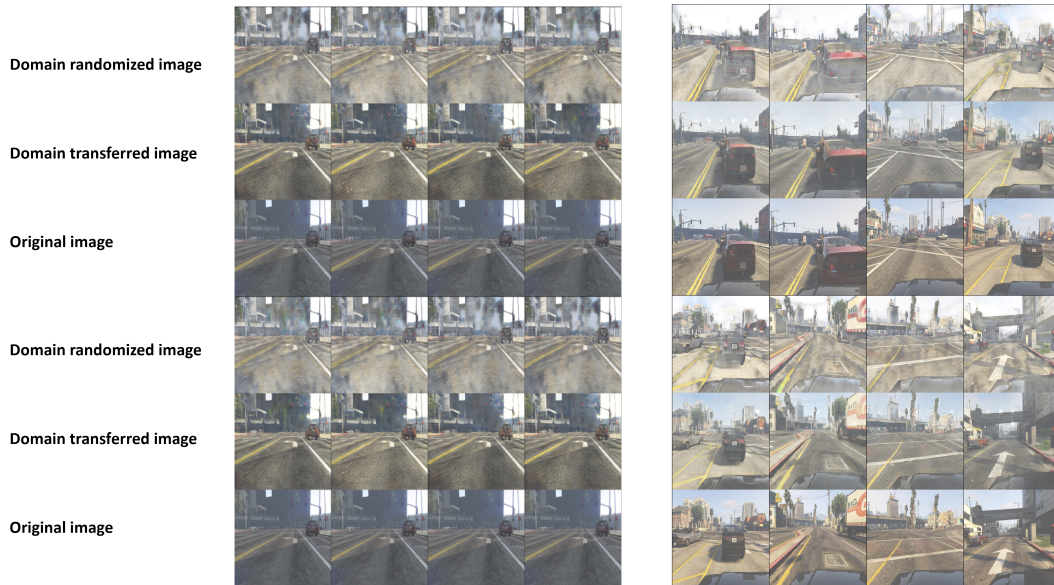


Figure 8. Comparison of the domain randomized, domain transferred and original images.

Notice that models trained under this assumption might perform poorly in real-world scenarios that diverge geometrically from the simulation. Further improving the model performance to overcome such cases can be considered to increase the randomization noise. To maximize randomization, we leveraged the advantage of single image to single image transfer, thereby increasing the variety of style images. This approach, detailed in Table 1, effectively balances the trade-off between randomization degree and geometric consistency.

4.3.3. Comparison to adaptation techniques

Our adaptation process requires no target domain data, relying instead on the prior knowledge that the simulation and real-world environments are similar geometrically but differ in texture and lighting, hence more challenging. This set the major difference between the proposed method and the existing domain adaptation strategies [22, 55]. As shown in Table 1, the model's generalization ability reaches a similar level of performance on the road, sidewalk, building, terrain and vegetation. The proposed strategy improves the classification on poles by 5–10 in general domain transfer tasks. However, there are some existing deficiencies in the small objects such as person and car that the classification performance reduced by 7 in mIoU for overall domain transfer task. This is acceptable especially consider the proposed method does not require target domain data comparing to the benchmark strategies. Typically, access to target data aids policy transfer, as data can be augmented towards the target domain distribution. However, in our experiments, using target domain data as noise does not improve the transfer process. This is because the learned intermediate feature representation of real-world data is relatively stable, and stylized images generated from real-world data do not provide sufficient randomization, leading to poor transfer performance. As none of the target domain images are used in our training process, the final model, which maps randomized to original data, is expected to generalize effectively across various real-world driving scenario datasets.

5. Conclusion and future work

In this paper, we introduce a comprehensive framework that integrates the benefits of both metric-based domain adaptation and traditional domain adaptation. This framework effectively utilizes perturbation injection, adding randomness to limited target domain data to better approximate the target distribution. We propose a domain generalization model using a triplet network, designed to address domain adaptation challenges in the absence of target domain data. This model employs contrastive learning to fine-tune random augmentations. Our approach demonstrates strong generalization capabilities, as evidenced by training a single model and successfully testing it on various unseen target domain datasets. However, a balance between the realism and randomization is needed as too much randomization potentially leading the network to learn irrelevant or detrimental features whereas insufficient randomization might not provide enough challenge to the model, failing to improve its generalization capabilities. Future work will explore the applicability of our framework in different domain adaptation tasks and investigate potential synergies with GAN-based transfer learning methods.

Conflicts of interests

The authors declare no conflict of interest.

Authors' contribution

Conceptualization, Chen Sun and Xingxin Chen; methodology, Chen Sun and Xingxin Chen; software, Xingxin Chen; validation, Xingxin Chen and Yukun Lu; formal analysis, Chen Sun; investigation, Yukun Lu; resources, Chen Sun; data curation, Xingxin Chen; writing—original draft preparation, Chen Sun and Xingxin Chen; writing—review and editing, Yukun Lu; visualization, Xingxin Chen; supervision, Dongpu Cao; project administration, Amir Khajepour. All authors have read and agreed to the published version of the manuscript.

References

- [1] Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, *et al.* Covariate shift and local learning by distribution matching. In *Dataset Shift in Machine Learning*, Cambridge, MA, USA: Biologische Kybernetik, 2009, pp. 131–160.
- [2] Sun C, Zhang R, Lu Y, Cui Y, Deng Z, *et al.* Toward Ensuring Safety for Autonomous Driving Perception: Standardization Progress, Research Advances, and Perspectives. *IEEE Trans. Intell. Transp. Syst.* 2024, 25(5):3286–3304.
- [3] Sun C, Cui Y, Đào ND, Mehrizi RV, Pirani M, *et al.* Medium-Fidelity Evaluation and Modeling for Perception Systems of Intelligent and Connected Vehicles. *IEEE Trans. Intell. Veh.* 2024, 9(1):1711–1721.
- [4] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. *arXiv* 2014, arXiv:1409.7495.
- [5] Zhang L. Transfer Adaptation Learning: A Decade Survey. *arXiv* 2019, arXiv:1903.04687.
- [6] Jordan M, Mitchell T. Machine Learning: Trends, Perspectives, and Prospects. *Science* 2015, 349(6245):255–260.
- [7] Dosovitskiy A, Ros G, Codevilla F, Lopez A, Koltun V. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*. 13–15 November 2017, pp. 1–16.
- [8] Shah S, Dey D, Lovett C, Kapoor A. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, Zurich, Switzerland, 12–15 September 2017, pp. 621–635.
- [9] Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain Adaptation via Transfer Component Analysis. *IEEE Trans. Neural Netw.* 2011, 22(2):199–210.
- [10] Wang J, Chen Y, Hu L, Peng X, Yu PS. Stratified Transfer Learning for Cross-domain Activity Recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Athens, Greece, 19–23 March 2018, pp. 1–10.
- [11] Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv* 2014, arXiv:1412.3474.
- [12] Long M, Wang J. Learning Transferable Features with Deep Adaptation Networks. *arXiv* 2015, arXiv:1502.02791.
- [13] Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial Discriminative Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hong Kong, China, 21–26 July 2017, pp. 7167–7176.
- [14] Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, *et al.* CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 10–15 July 2018, pp. 1989–1998.
- [15] Chen X, Wang S, Long M, Wang J. Transferability vs. Discriminability: Batch Spectral Penalization for Adversarial Domain Adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, 9–15 June 2019, pp. 1081–1090.
- [16] Kar A, Prakash A, Liu MY, Cameracci E, Yuan J, *et al.* Meta-Sim: Learning to Generate Synthetic Datasets. *arXiv* 2019, arXiv:1904.11621.
- [17] Yang W, Wang X, Farhadi A, Gupta A, Mottaghi R. Visual Semantic Navigation using Scene Priors. *arXiv* 2018, arXiv:1810.06543.
- [18] Finn C, Abbeel P, Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv* 2017, arXiv:1703.03400.
- [19] Yue X, Zhang Y, Zhao S, Sangiovanni-Vincentelli A, Keutzer K, *et al.* Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data. *arXiv* 2019, arXiv:1909.00889.
- [20] Long M, Cao Z, Wang J, Jordan MI. Conditional Adversarial Domain Adaptation. *arXiv* 2018, arXiv:1705.10667.
- [21] French G, Mackiewicz M, Fisher M. Self-ensembling for visual domain adaptation. *arXiv* 2018, arXiv:1706.05208.
- [22] Hoffman J, Wang D, Yu F, Darrell T. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. *arXiv* 2016, arXiv:1612.02649.
- [23] Ruiz N, Schuler S, Chandraker M. Learning To Simulate. *arXiv* 2019, arXiv:1810.02513.
- [24] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, MIT Press, 2006, pp. 137–144.

- [25] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A Kernel Two-Sample Test. *J. Mach. Learn. Res.* 2012, 13:723–773.
- [26] Liang J, Hu D, Feng J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine*. 13–18 July 2020, vol. 119, pp. 6028–6039.
- [27] Xia H, Zhao H, Ding Z. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11–17 October 2021, pp. 9010–9019.
- [28] Yeh HW, Yang B, Yuen PC, Harada T. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5–9 January 2021, pp. 474–483.
- [29] Yang S, Wang Y, Van De Weijer J, Herranz L, Jui S. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11–17 October 2021, pp. 8978–8987.
- [30] Li J, Yu Z, Du Z, Zhu L, Shen HT. A comprehensive survey on source-free domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2024, 46(8):1–22.
- [31] Pashevich A, Strudel RAM, Kalevatykh I, Laptev I, Schmid C. Learning to Augment Synthetic Images for Sim2Real Policy Transfer. *arXiv* 2019, arXiv:1903.07740.
- [32] Zoph B, Cubuk ED, Ghiasi G, Lin TY, Shlens J, *et al.* Learning Data Augmentation Strategies for Object Detection. *arXiv* 2019, arXiv:1906.11172.
- [33] Cubuk ED, Zoph B, Shlens J, Le QV. RandAugment: Practical automated data augmentation with a reduced search space. *arXiv* 2019, arXiv:1909.13719.
- [34] Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv* 2017, arXiv:1703.10593.
- [35] James S, Wohlhart P, Kalakrishnan M, Kalashnikov D, Irpan A, *et al.* Sim-to-Real via Sim-to-Sim: Data-efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks. *arXiv* 2018, arXiv:1812.07252.
- [36] Mirza M, Osindero S. Conditional Generative Adversarial Nets. *arXiv* 2014, arXiv:1411.1784.
- [37] Huang X, Belongie S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *arXiv* 2017, arXiv:1703.06868.
- [38] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv* 2018, arXiv:1801.03924.
- [39] Wang J, Song Y, Leung T, Rosenberg C, Wang J, *et al.* Learning Fine-grained Image Similarity with Deep Ranking. *arXiv* 2014, arXiv:1404.4661.
- [40] Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, 13–15 May 2010, pp. 297–304.
- [41] Graf F, Hofer C, Niethammer M, Kwitt R. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*. 18–24 July 2021, pp. 3821–3830.
- [42] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* 2015, arXiv:1411.4038.
- [43] Richter SR, Vineet V, Roth S, Koltun V. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 11–14 October 2016, pp. 102–118.
- [44] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, *et al.* The Cityscapes Dataset for Semantic Urban Scene Understanding. *arXiv* 2016, arXiv:1604.01685.
- [45] Yu F, Xian W, Chen Y, Liu F, Liao M, *et al.* BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. *arXiv* 2018, arXiv:1805.04687.
- [46] Sankaranarayanan S, Balaji Y, Jain A, Lim SN, Chellappa R. Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation. *arXiv* 2018, arXiv:1711.06969.
- [47] Zhang Y, David P, Gong B. Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. *arXiv* 2017, arXiv:1707.09465.
- [48] Saito K, Watanabe K, Ushiku Y, Harada T. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. *arXiv* 2017, arXiv:1712.02560.
- [49] Huang J, Guan D, Xiao A, Lu S. FSDR: Frequency Space Domain Randomization for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer*

- Vision and Pattern Recognition (CVPR)*. 19–25 June 2021, pp. 6891–6902.
- [50] Kim S, Kim Dh, Kim H. Texture learning domain randomization for domain generalized segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 2–6 October 2023, pp. 677–687.
 - [51] Zhang Y, David P, Foroosh H, Gong B. A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* 2019, 42(5):1823–1841.
 - [52] Qi R, Rushton M, Khajepour A, Melek WW. Decoupled modeling and model predictive control of a hybrid cable-driven robot (HCDR). *Robot. Auton. Syst.* 2019, 118:1–12.
 - [53] Sun C, Wang C, Deng Z, Cao D. Dimensionless Model-Based System Tracking Via Augmented Kalman Filter for Multiscale Unmanned Ground Vehicles. *IEEE/ASME Trans. Mechatron.* 2021, 26(2):600–610.
 - [54] Sun C, Cui Y, Lu Y, Cao Y, Cao D, *et al.* Robust Learning for Autonomous Driving Perception Tasks in Cyber-Physical-Social Systems. In *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, Orlando, FL, USA, 7–9 November 2023, pp. 1–7.
 - [55] Chen Z, Zhuang J, Liang X, Lin L. Blending-target Domain Adaptation by Adversarial Meta-Adaptation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 16–20 June 2019, pp. 297–304.