GCISG: Guided Causal Invariant Learning for Improved Syn-to-real Generalization

Gilhyun Nam[1], Gyeongjae Choi[1], and Kyungmin Lee[2]

[1] Agency for Defense Development (ADD), Daejeon, Korea
{ngh707, def6488}@gmail.com
[2] Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
kyungmnlee@kaist.ac.kr

Abstract. Training a deep learning model with artificially generated data can be an alternative when training data are scarce, yet it suffers from poor generalization performance due to a large domain gap. In this paper, we characterize the domain gap by using a causal framework for data generation. We assume that the real and synthetic data have common content variables but different style variables. Thus, a model trained on synthetic dataset might have poor generalization as the model learns the nuisance style variables. To that end, we propose causal invariance learning which encourages the model to learn a style-invariant representation that enhances the syn-to-real generalization. Furthermore, we propose a simple yet effective feature distillation method that prevents catastrophic forgetting of semantic knowledge of the real domain. In sum, we refer to our method as Guided Causal Invariant Syn-to-real Generalization that effectively improves the performance of syn-to-real generalization. We empirically verify the validity of proposed methods, and especially, our method achieves state-of-the-art on visual syn-to-real domain generalization tasks such as image classification and semantic segmentation.

1 Introduction

While deep neural networks have shown their great capability on various computer vision tasks, the majority of them count on a sufficient amount of training data with qualified labels. However, obtaining data or labels is expensive or difficult. Thus, a manually generated training dataset can be an alternative to the shortage or absence of training data. Also, by using the computer graphics engine [18,30,32], one can obtain labels without any cost of human labor [16,39].

However, training with synthetic data cannot fully replace the real training data as they suffer from poor generalization performance in the real domain. Many previous works explain the reason for inferior performance by the existence of a large domain gap [29,42] which makes the model overfits to the synthetic domain. In this paper, we break down the domain gap between the synthetic and real data by using the structural causal model. We build a causal model that an image is generated from two latent variables: style variables such as texture and content variables such as shape. We assume that the synthetic and real data differ in style variables while having common content variables that are relevant to the downstream tasks. Furthermore, we also assume that the model learns the task-irrelevant style variables, therefore exhibiting poor generalization on the real domain. The recent studies [1,17] show that convolutional neural networks tend to rely on texture rather than shape which supports our claim.

To that end, we propose causal invariant learning for syn-to-real generalization by promoting causal invariance loss to learn style-invariant representation. Our method is related to the contrastive learning method, which learns representation by aligning positive pairs. We apply data augmentation to generate a pair of images with the same content but different style variables, then align their distributions of representations to achieve style-invariance. Our loss function is similar to that of [50,15,46,34], where those methods are for other than syn-to-real generalization such as self-supervised learning or knowledge distillation.

Furthermore, many methods show that using real-domain guidance with ImageNet pre-trained model for regularization is effective for syn-to-real generalization [6,5]. Chen et al. [6] utilized the knowledge distillation loss by re-using the pre-trained ImageNet classifier and Chen et al. [5] used intermediate features of the ImageNet pre-trained network to compute negative samples in contrastive learning. Those methods posit that the ImageNet classification is strongly correlated to the syn-to-real generalization performance. Instead, we propose to directly guide the model by simply regularizing the feature distance with the ImageNet pre-trained model, yet the key is to extract the semantic knowledge by using self-attention pooling. We empirically show that the simple guidance loss with self-attention pooling successfully guides the model to localize the object in the image, and helps the syn-to-real generalization.

In summary, we propose *Guided Causal Invariant Syn-to-real Generalization* (GCISG), where we propose causal invariance loss that helps to learn style-invariant features and guidance loss with self-attention pooling that helps extract semantic information from ImageNet pre-trained neural network. Our contributions are listed below:

- By adopting the causal framework for a syn-to-real setup, we propose causal invariance loss which regularizes the model to learn a style-invariant representation that is relevant to syn-to-real generalization.
- We present a simple and effective guidance loss that guides the model with real-domain guidance from ImageNet pre-trained model by utilizing self-attention pooling.
- We empirically show that our method significantly outperforms competing syn-to-real generalization methods on various vision tasks such as image classification, semantic segmentation, and object detection.

2 Related Work

Domain generalization. The domain generalization problem aims to train on the source domain and generalize to an unseen target domain. The biggest

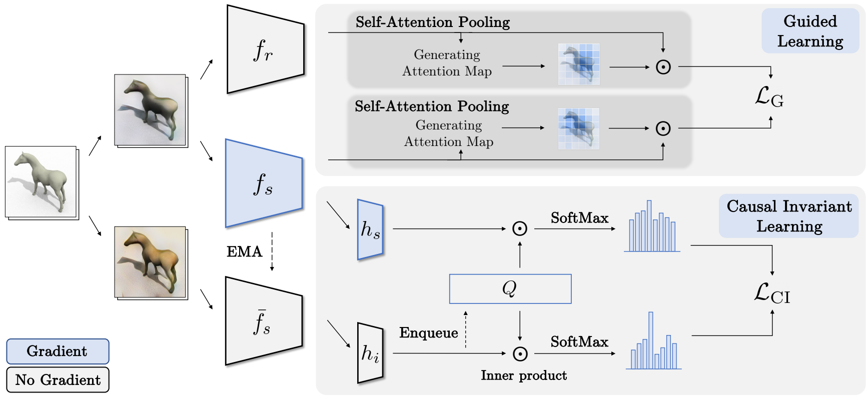

Fig. 1. Overall demonstration of proposed GCISG. There are two branches: 1) the guidance loss that preserves semantic information from ImageNet pre-trained model by using self-attention pooling, and 2) the causal invariance loss that learns style-invariant representation using contrastive learning framework.

concern of domain generalization is that there is a large domain gap between source and target domains, which deteriorates the performance. Therefore, many methods proposed an adversarial learning framework to match the feature distributions with a prior distribution [24] or diversify the style of the source domain to enhance the generalization [47,44]. Others focused on the effect of batch normalization on domain generalization. Pan et al. [33] showed that using instance normalization can boost the generalization, and RobustNet [10] used instance selective whitening to improve the robustness. Recent studies showed that using the knowledge from the real domain is effective for syn-to-real domain generalization problems [5,6,9], and our work is concurrent to those approaches.

Contrastive learning and style invariance. Recently, contrastive learning with multi-view data augmentation has shown their efficacy in self-supervised representation learning [45,19,4,31] and domain generalization [28]. Our contrastive learning objective is similar to that of SEED [15], ReSSL [50], and RELIC [31], while SEED uses the contrastive objective to distill the representational knowledge on smaller networks, and ReSSL and RELIC are for self-supervised representation learning. Remark that CSG [5] also uses contrastive learning for syn-to-real generalization, while it is different from ours as they compute contrastive loss across ImageNet pre-trained networks, while ours aim to learn style-invariant representation on the synthetic data.

Learning with guidance. Training a model by guidance with a pre-trained model's knowledge is not only effective for syn-to-real generalization but also for various machine learning tasks such as knowledge distillation [21,43] and incremental learning [25,2]. In syn-to-real generalization or adaptation, models are enforced to retain the knowledge from the ImageNet pre-trained networks by minimizing the feature distance [9], matching the outputs of ImageNet classification [6], or using the contrastive learning objective [5]. Similarly, the knowledge distillation methods use similar tactics to compress the knowledge from bigger networks to smaller ones [41,46]. And for incremental learning, similar methods were used to prevent catastrophic forgetting [36].

Our method uses self-attention pooled feature minimization to guide the model to enhance the generalization on a real domain. Transferring the knowledge by using attention has been also studied for knowledge distillation [48], and incremental learning [13]. While Chen et al. [5] used attention pooling for syn-to-real generalization, there are differences in that they pool features for contrastive learning, while our objective directly minimizes the distance between them.

3 Proposed Method

3.1 Causal Invariance for Syn-to-real Generalization

We consider a causal model for synthetic and real data generation. In Fig. 2, we assume that the synthetic data $X_{\texttt{syn}}$ is generated from the content variable C and style variable $S_{\texttt{syn}}$, and a real data $X_{\texttt{real}}$ is generated from the content variable C and style variable $S_{\texttt{real}}$. Thus, we claim that the domain gap occurs because of the difference between the latent style variables of the two domains. We also assume that the task label Y is only dependent on the content variable. Those assumptions give us insight into why the model shows inferior generalization when solely trained with synthetic data: the model learns nuisance style variables irrelevant to the task. Also, a previous study [17] corroborates our hypothesis that the convolutional neural networks are biased toward texture (i.e. the style variable).

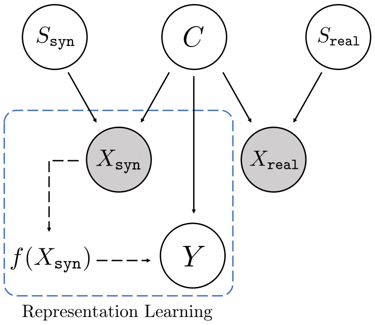

Fig. 2. Causal graph for synthetic and real data.

If one can extract style-invariant or content-preserving representation from the synthetic data, the model can enhance its generalization capability in the real domain. Thus, our goal is to extract style-invariant representation which is useful in the generalization on the real domain. To that end, we propose causal invariant syn-to-real generalization which is composed of two folds: first, we further diversify the style variable by using a strong data augmentation module such as RandAugment [12], and second, we use a contrastive learning framework to learn style-invariant representations.

For the synthetic training model f_s, we attach projector h_s to project onto a smaller dimension. For projector, we use two-layer multi-layer perceptron (MLP) as used in many contrastive representation learning frameworks [4,7]. Let us denote the encoder by $g_s = h_s \circ f_s$ and we build target encoder $\bar{g}_s = h_i \circ \bar{f}_s$, while h_i is the initial state of projector h_s. The weights of $\bar{f}_s$ are updated by the exponential moving average of the weight of the encoder f_s. Given a synthetic data x, let x_1 and x_2 be the two augmented views of x. Denote $z_1 = g_s(x_1)/\|g_s(x_1)\|_2 \in \mathbb{R}^d$ and $\bar{z}_2 = \bar{g}_s(x_2)/\|\bar{g}_s(x_2)\|_2 \in \mathbb{R}^d$ be the outputs of each encoder and target encoder with ℓ_2 normalization. To achieve stable causal invariance, we aim to match the distributions of z_1 and $\bar{z}_2$ with enough amount of support samples, which is a similar approach to prior works in other tasks [15,31,50]. We pertain support samples $Q \in \mathbb{R}^{K\times d}$, and aim to minimize the probabilistic discrepancy between the relational distributions $p(z_1; Q)$ and $p(\bar{z}_2; Q)$.

$$p_\tau(z; Q)[k] \coloneqq \frac{\exp\left(z^\top q_k/\tau\right)}{\sum_{k'=1}^{K} \exp\left(z^\top q_{k'}/\tau\right)}, \tag{1}$$

where q_k is the k-th component of support samples Q and $\tau > 0$ is a temperature hyperparameter. Then we define the causal invariance loss between z_1 and $\bar{z}_2$ by the cross-entropy loss between the relation similarities as following:

$$\ell_{\texttt{CI}}(z_1, \bar{z}_2) = -\sum_{k=1}^{K} p_{\bar{\tau}}(\bar{z}_2; Q)[k] \log\left(p_\tau(z_1; Q)[k]\right), \tag{2}$$

where we use distinct $\tau, \bar{\tau} > 0$ to regulate the sharpness of distributions. For computational efficiency, we symmetrically compute compute between $\bar{z}_1 = \bar{g}_s(x_1)/\|\bar{g}_s(x_1)\|_2$ and $z_2 = g_s(x_2)/\|g_s(x_2)\|_2$, and the total causal invariance loss for a data x is given by

$$\mathcal{L}_{\texttt{CI}}(x) = \frac{1}{2}\ell_{\texttt{CI}}(z_1, \bar{z}_2) + \frac{1}{2}\ell_{\texttt{CI}}(z_2, \bar{z}_1). \tag{3}$$

The support samples Q are managed by a queue, where we enqueue the outputs of the target encoder at each iteration, and dequeue the oldest ones. We set K to 65536 to ensure the relational distribution contains sufficient information to learn style-invariant representations. Remark that the temperature parameter greatly affects the stability of the learning with causal invariance loss. Generally, we choose $\bar{\tau}$ to be smaller than τ, so that the target distribution is sharper.

Dense causal invariant learning for semantic segmentation. For the semantic segmentation task, an image contains various objects and semantic features. Therefore, we use dense causal invariance loss where we compute the loss over the patches of the representations. Following the method in [5], we crop the feature maps into N_l patches and pass them forward to the projector to compute causal invariance loss on each cropped representation:

$$\mathcal{L}_{\texttt{CI-Dense}}(x) = \frac{1}{N_l}\sum_{n=1}^{N_l} \frac{1}{2}\ell_{\texttt{CI}}\left(z_{1,n}, \bar{z}_{2,n}\right) + \frac{1}{2}\ell_{\texttt{CI}}\left(z_{2,n}, \bar{z}_{1,n}\right), \tag{4}$$

where each $z_{i,n}, \bar{z}_{i,n}$ is the encoder output of n-th feature map. In our experiments, we use $N_l = 8 \times 8$.

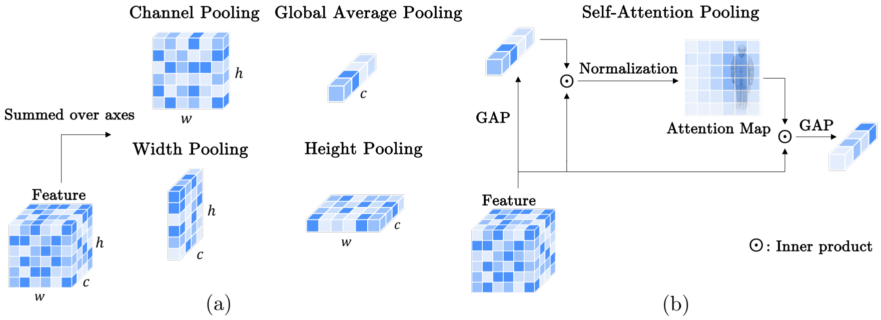

Fig. 3. (a) Example images of channel, global average, width and height pooled feature map. (b) Mechanism of self-attention pooling for guidance loss.

3.2 Guided Learning for Syn-to-real Generalization

In this section, we present a simple, yet effective method to guide a model with real-domain guidance from the ImageNet-pretrained model by using feature distillation. In contrast to previous approaches such as ASG [6] and CSG [5], our method does not require any task-specific information or complicated loss function.

Pooled feature distillation. For a data x, let $f_s(x), f_r(x) \in \mathbb{R}^{C\times H\times W}$ be the feature map of an intermediate convolutional layer of ImageNet pre-trained network f_r and training network f_s. Then it is straightforward to minimize the squared distance between the feature maps $f_s(x)$ and $f_r(x)$ for guidance loss to retain the knowledge of f_r as much as possible while learning task-related knowledge from the synthetic data. To design the guidance loss, there are two opposite concepts that must be considered: rigidity and plasticity. The rigidity is on how much the model can retain the knowledge of f_r, and the plasticity accounts for the flexibility of the model to learn from the synthetic domain.

While the direct minimization with the feature maps has high rigidity, it has low plasticity that the model cannot sufficiently learn task-related knowledge with synthetic data. Douillard et al. [14] showed that by using an appropriate pooling operator, the model can balance the rigidity and plasticity to conduct incremental learning. Similarly, we use the pooling operator to extract sufficient information from the feature maps and balance the rigidity and plasticity to improve syn-to-real generalization.

In Fig. 3 (a), we list various pooling methods that we consider in this paper. Let P be such pooling operators in Fig. 3 (a), then the guidance loss is given by the ℓ_2 distance between the normalized pooled feature maps of $f_s(x)$ and $f_r(x)$:

$$\mathcal{L}_{\mathtt{G}}(x;P) = \left\| \frac{P(f_s(x))}{\|P(f_s(x))\|_2} - \frac{P(f_r(x))}{\|P(f_r(x)\|_2} \right\|_2^2 \tag{5}$$

Improved guidance by self-attention pooling. Alternatively, we propose self-attention pooling based guidance loss which improves syn-to-real generalization. The self-attention pooling captures semantically important features by multiplying the importance weight on each feature map. Therefore, the guidance with self-attention pooling allows the model to learn the semantically important feature that f_r focuses on. The self-attention pooling is demonstrated in Fig. 3 (b). Let $v \in \mathbb{R}^{C\times W\times H}$ be a feature map, then the attention map $a \in \mathbb{R}^{W\times H}$ is computed by the normalization on the global average pooling of v:

$$a[w,h] = \frac{\sum_{c=1}^{C} v[c,w,h]g[c]}{\sum_{w'=1}^{W}\sum_{h'=1}^{H}\sum_{c=1}^{C} v[c,w',h']g[c]}, \tag{6}$$

where $g[c] = \frac{1}{HW}\sum_{w=1}^{W}\sum_{h=1}^{H} v[c,w,h]$ is global average pooling of v. Then the self-attention pooling operator P_a outputs an importance-weighted feature map with weights given by the attention map:

$$P_a(v)[c] = \sum_{w=1}^{W}\sum_{h=1}^{H} v[c,w,h]a[w,h]. \tag{7}$$

Then our final guidance loss with self-attention pooling operator P_a is given by

$$\mathcal{L}_{\mathtt{G}}(x;P_a) = \left\| \frac{P_a(f_s(x))}{\|P_a(f_s(x))\|_2} - \frac{P_a(f_r(x))}{\|P_a(f_r(x)\|_2} \right\|_2^2 \tag{8}$$

The empirical analysis on the effect of pooling on the guidance loss is explored in section 4.2.

3.3 Guided Causal Invariant Syn-to-real Generalization

In this section, we present the overall description of our method for syn-to-real generalization, which we refer to *Guided Causal Invariant Syn-to-real generalization* (GCISG). Given a task loss $\mathcal{L}_{\mathtt{task}}$ that takes synthetic data x with corresponding label y, the GCISG adds guidance loss $\mathcal{L}_G$ that regularizes the model to retain the semantic knowledge of ImageNet pre-trained model and causal invariance loss $\mathcal{L}_{\mathtt{CI}}$ to achieve style-invariant representation that promotes generalization on a real domain. Thus, the overall loss is computed by following:

$$\mathcal{L}(x) = \mathcal{L}_{\mathtt{Task}}(x,y) + \lambda_{\mathtt{G}}\mathcal{L}_{\mathtt{G}}(x) + \lambda_{\mathtt{CI}}\mathcal{L}_{\mathtt{CI}}(x). \tag{9}$$

Remark that our GCISG framework is agnostic to tasks, as guidance loss and causal invariance loss are computed in an unsupervised manner.

Stage-wise loss computation. The causal invariance loss $\mathcal{L}_{\texttt{CI}}$ and the guidance loss $\mathcal{L}_{\texttt{G}}$ can be computed for each intermediate layer. Denote $f^{(l)}$ to be l-th layer of the neural network, (e.g. the block layers of ResNet [20]), then the GCISG is computed for each layer by following:

$$\mathcal{L}(x) = \mathcal{L}_{\texttt{Task}}(x, y) + \sum_{l} \lambda_{\texttt{G}}^{(l)} \mathcal{L}_{\texttt{G}}^{(l)}(x) + \sum_{l} \lambda_{\texttt{CI}}^{(l)} \mathcal{L}_{\texttt{CI}}^{(l)}(x). \tag{10}$$

In section 4.2, we conduct an ablation study on the effect of choosing layer.

4 Experiments

In this section, we empirically validate the effectiveness of our method on various syn-to-real generalization tasks. We report the performance of a synthetic-trained model in the unseen real domain for evaluation. Furthermore, we analyze the model with various auxiliary evaluation metrics to support our claims.

Evaluation metrics for causal invariance. To quantify the style-invariance of a representation, we introduce *match rate* ($\mathcal{M}$), which evaluates the consistency of a model on images that have the same semantics but different styles. For each image, we generate another stylized image and check if it produces the same output as the original one. Formally, we report the match rate by the number of consistent samples out of the number of the entire validation set. To generate stylized images, we use photometric transforms such as Gaussian blurring and color jittering as done in [10].

Evaluation metrics for guidance. We present two different measures to quantify the quality of guidance from the ImageNet pre-trained model. First, we bring the linear classification layer of the pre-trained ImageNet model and attach it to the synthetic-trained model. Then we report the ImageNet validation accuracy ($\text{Acc}_{\texttt{IN}}$) by inferring on the ImageNet validation dataset. Second, we evaluate the similarity between ImageNet pre-trained model and the synthetic-trained model by using *centered kernel alignment* (CKA) similarity [23]. Here, the high ImageNet validation accuracy demonstrates that our model avoids the catastrophic forgetting of the task information, and the high CKA similarity proves the preservation of representational knowledge of a real pre-trained model. For the computation of CKA similarity, we collect the features of the penultimate layers of the ImageNet pre-trained model and synthetic-trained model on the validation dataset. The detailed implementations are in the supplementary material.

4.1 Image Classification

Datasets. We demonstrate the effectiveness of GCISG on VisDA-17 [35] image classification benchmark. The synthetic training set consists of the images rendered from 3D models with various angles and illuminations. The real validation

Table 1. Top-1 accuracy (%) on VisDA-17 (`VisDA`) and ImageNet (`IN`) validation datasets of various syn-to-real generalization methods. All methods used ResNet-101 as base architecture.

Method	$\text{Acc}_{\texttt{VisDA}}$	$\text{Acc}_{\texttt{IN}}$
Oracle on ImageNet	53.3	**77.4**
Vanilla L2 distance	56.4	49.1
ROAD [49]	57.1	**77.4**
SI [9]	57.6	53.9
ASG [6]	61.1	76.7
CSG [5]	64.1	73.8
GCISG	**67.5**	75.4

Table 2. Top: Comparison with CSG and oracle on ImageNet by evaluation metrics for invariance ($\mathcal{M}$) and guidance ($\text{Acc}_{\texttt{IN}}$, CKA). The oracle freezes the backbone and fine-tune the classification head. Bottom: Effects of $\mathcal{L}_{\texttt{G}}$ and $\mathcal{L}_{\texttt{CI}}$ in GCISG.

Method	$\mathcal{L}_{\texttt{G}}$	$\mathcal{L}_{\texttt{CI}}$	$\text{Acc}_{\texttt{VisDA}}$	$\mathcal{M}$	$\text{Acc}_{\texttt{IN}}$	CKA
Oracle	✗	✗	53.3	69.5	77.4	1.0
CSG [5]	✓	✗	64.1	77.0	73.8	0.89
GCISG	✗	✗	55.9	73.3	52.7	0.72
	✗	✓	58.6	74.4	51.9	0.73
	✓	✗	64.3	76.6	**75.9**	**0.95**
	✓	✓	**67.5**	**78.6**	75.4	0.93

set is a subset of MS-COCO [27]. Each dataset contains the same 12 object categories.

Implementation details. We adopt ImageNet pre-trained ResNet-101 as the backbone. We use an SGD optimizer with a learning rate of 0.001, a batch size of 64 for 30 epochs. For data augmentations, we use the RandAugment [12] following the protocol of CSG [5]. The momentum rate is 0.996 and the number of the support sample is 65536 throughout the training.

Results. In Table 1, we report the top-1 validation accuracy on VisDA-17 ($\text{Acc}_{\texttt{VisDA}}$) of synthetic-trained model with various syn-to-real generalization methods. We also report the ImageNet validation accuracy ($\text{Acc}_{\texttt{IN}}$) of the same synthetic-trained model as a guidance metric. We observe that GCISG outperforms the competing methods on VisDA-17 on a large margin. As shown in Table 1, there is a correlation between the ImageNet classification accuracy and syn-to-real generalization performance. However, the correlation is not strong that the methods with higher ImageNet accuracy such as ASG [6] and ROAD [9], do not necessarily have higher validation accuracy on VisDA-17. Thus, sufficiently high ImageNet validation accuracy might be useful for syn-to-real generalization, but it may hamper the generalization if it is too much as it transfers knowledge that is overfitted to the ImageNet classification task.

Table 2 shows that our method not only achieves a higher VisDA-17 validation accuracy but also has a higher match rate ($\mathcal{M}$) and CKA similarity than CSG. The higher match rate indicates that the learned representation is invariant to the style change, and the higher CKA similarity indicates the conformation of learned representation to the ImageNet pre-trained network. Therefore, one can observe that both factors have positive impacts on syn-to-real generalization.

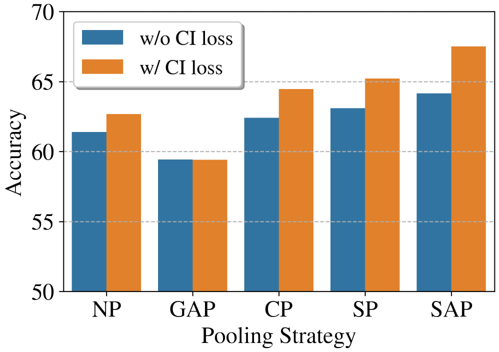

Fig. 4. Ablation of feature pooling strategies for $\mathcal{L}_{\mathrm{G}}$. NP, GAP, CP, SP, and SAP denote no pooling, global average pooling, channel pooling, spatial pooling, and self-attention pooling respectively.

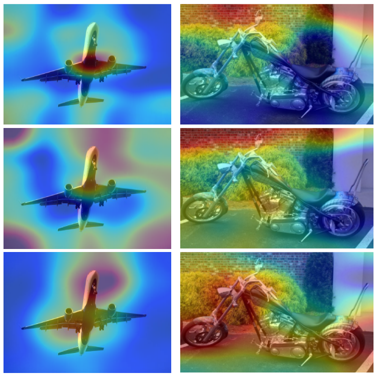

Fig. 5. Visualized attention of feature pooling methods. From top to bottom, rows correspond to NP, GAP, SAP. The red area corresponds to high score for class.

Table 2 reports the effect of using causal invariance loss and guidance loss on VisDA-17 validation accuracy and auxiliary evaluation metrics. Remark that CSG [5] applied guidance loss in contrastive learning with ImageNet pre-trained features. We observe that the guidance loss generally boosts ImageNet validation accuracy and CKA similarity. This clearly demonstrates that the transferring knowledge from the ImageNet pre-trained model which holds useful knowledge of real domain helps the syn-to-real generalization. Moreover, using the causal invariance loss generally boosts both match rate and VisDA-17 validation accuracy. This also demonstrates that the style-invariance of the model enhances the syn-to-real generalization.

4.2 Ablation Study

Ablation on pooling operators for guidance loss. As explained in section 3.2, the choice of pooling operator for guidance loss affects the learning. We conduct an ablation study on the effect of the pooling operator on VisDA-17 syn-to-real generalization. We compare 5 different pooling strategies: no pooling (NP), global average pooling (GAP), channel pooling (CP), spatial pooling (SP), and self-attention pooling (SAP), where spatial pooling takes both height pooling and width pooling. In Fig. 4, we show the generalization performance in VisDA-17 for various pooling strategies. We also conduct the ablation study on removing causal invariance loss to identify the effect of the pooling operator on the generalization. In both situations, we observe that self-attention pooling outperforms other pooling strategies.

We also investigate the attention map using Grad-CAM [40] to qualitatively analyze the effect of pooling operators for guidance loss. In Fig. 5, we observe a tendency that SAP comprehensively includes more semantically meaningful

Table 3. Ablation on the choice of feature stages S_{G} and S_{CI}.

S_{G}	S_{CI}	Acc
{0,1,2}	{0,1,2}	59.8
{0,1,2}	{3,4}	55.9
{3,4}	{0,1,2}	66.2
{3,4}	{3,4}	**67.5**

Table 4. Ablation on the choice of temperatures $\tau, \bar{\tau}$ for $\mathcal{L}_{\mathrm{CI}}$.

$\bar{\tau}$	0.02	0.04	0.06	0.08
$\tau = 0.12$	66.6	**67.5**	66.5	66.3
τ	0.08	0.1	0.12	0.14
$\bar{\tau} = 0.04$	66.8	67.1	**67.5**	67.4

Table 5. Ablation on the loss functions for style-invariant learning.

Loss	$\mathcal{L}_{\mathrm{G}}$	Acc
InfoNCE	✗	56.2
	✓	64.7
CI loss	✗	58.6
	✓	**67.5**

regions than NP which only focuses on detailed features with relatively narrow receptive fields. Also, one can observe that SAP focuses less on meaningless parts compared to GAP, which does not sufficiently encode informative features with a relatively wide receptive field.

Ablation on feature stages. We conduct an ablation study on generalization performance with different feature stage for $\mathcal{L}_{\mathrm{CI}}$ and $\mathcal{L}_{\mathrm{G}}$. Let us denote S_{G} and S_{CI} to be the set of layers that we used for computation of guidance loss and causal invariance loss. In ResNet-101, there is one convolutional layer and 4 Res-Blocks. We grouped the first three layers and last two layers and experimented on 4 combinations of guidance loss and causal invariance loss with different layer groups. As shown in Table 3, the computation of guidance loss and causal invariance loss with deeper feature stages is more effective in generalization than the shallower one, as the deeper stage of features holds richer semantic information. This result is consistent with the tendency of deep layers to encode content discriminative information as Pan et al. pointed out in IBN-Net [33].

Ablation on temperature scale. We conduct an ablation study on the effect of different temperature scales for causal invariance loss. In Table 4, when $\tau = 0.12$ we observe that small value of $\bar{\tau}$ is beneficial for the performance. Conversely, when $\bar{\tau}$ is fixed to 0.04, relatively high value of τ has better performance. Thus, we choose $\tau = 0.12$ and $\bar{\tau} = 0.04$ for our experiments.

Ablation on causal invariance loss. We conduct an ablation study on the effect of different contrastive losses for causal invariance loss. We compare proposed $\mathcal{L}_{\mathrm{CI}}$ and InfoNCE loss [7] on VisDA-17 image classification. To see the effect of causal invariance loss solely, we further removed the guidance loss. In Table 5, we observe that $\mathcal{L}_{\mathrm{CI}}$ is more effective in syn-to-real-generalization that InfoNCE loss regardless of the guidance loss.

4.3 Semantic Segmentation

Datasets. For semantic segmentation, the source domain is GTAV [38] and the target domain is Cityscapes [11]. GTAV is a large-scale dataset containing

Table 6. Comparison of mIoU (%) with existing synthetic-to-real generalization and domain generalization methods on GTAV-Cityscapes datasets. All methods used ResNet-50 as base architecture.

Method	IBN-Net [33]	DRPC [47]	RobustNet [10]	ASG [6]	CSG [5]	**GCISG**
No adapt	22.17	32.45	28.95	23.29	25.43	26.67
Adapt	29.64	37.42	36.58	29.65	35.27	**39.01**
mIoU ↑(%)	7.47	4.97	7.63	6.36	9.84	**12.34**

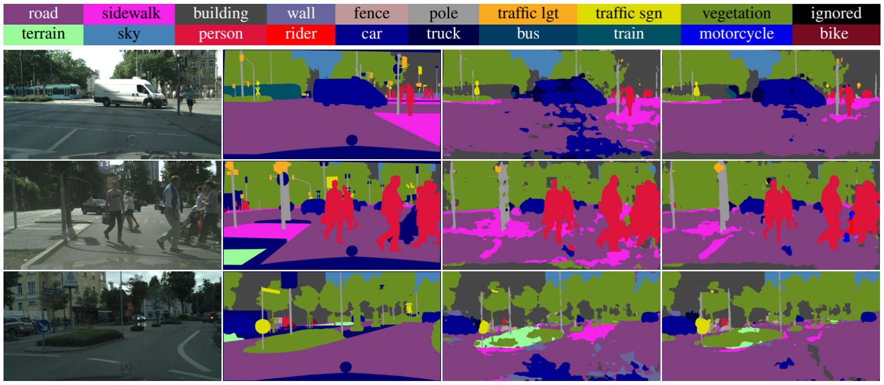

Fig. 6. Segmentation results on GTAV to Cityscapes. From left to right, columns correspond to original images, ground truth, prediction results of CSG, and GCISG.

24,996 synthetic driving-scene images with resolution 1914×1052, taken from the Grand Theft Auto V game engine. Cityscapes street scene dataset contains 2,975 train images and 500 validation images with resolution 2048×1024, taken from European cities. All the images in the source and target dataset have pixel-level annotations with 19 semantic categories. For evaluation, we report the mean intersection over union (mIoU) on the Cityscapes validation set.

Implementation details. We adopt DeepLab-V3 [3] with ImageNet pre-trained ResNet-50 backbone for our semantic segmentation network. We employ an SGD optimizer with a learning rate of 0.001 and a batch size of 8 for 40 epochs. We crop an image with the size of 512×512 and apply color jittering and multi-scale resizing.

Results. In Table 6, we report the mIoU (%) of the semantic segmentation models trained with various syn-to-real generalization methods. We observe that the model trained with GCISG achieves the best performance gain. It is worth noting that IBN-Net [33] and RobustNet [10] modifies the normalization blocks, and DRPC [47] requires preparation steps of stylizing images before training.

Table 7. The average precision (AP) of Sim10k to Cityscapes object detection with various methods. All experiments are run by us.

Method	AP	AP_{50}	AP_{75}
Baseline	24.8	45.1	24.0
CSG	28.5	50.4	28.1
GCISG	**30.6**	**54.6**	**29.5**

Table 8. Top-1 accuracy (%) of unsupervised domain adaptation on VisDA-17 by using CBST initialized with different methods.

Method	$\text{Acc}_{\texttt{VisDA}}$
Source only + CBST [52]	76.4
ASG + CBST [6]	82.5
GCISG + CBST	**83.6**

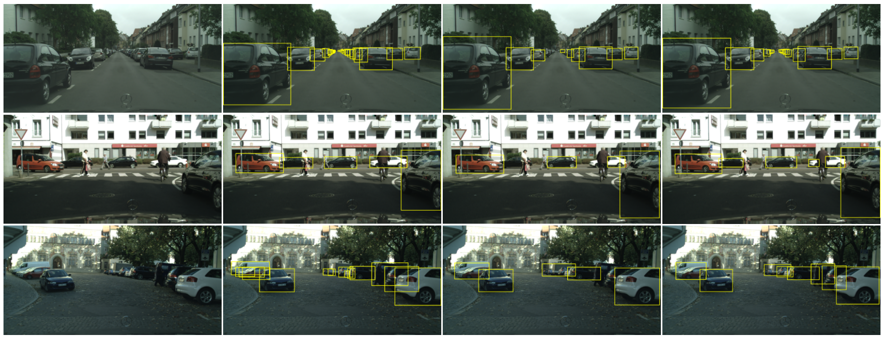

Fig. 7. Object detection results on Sim10k to Cityscapes. From left to right, columns correspond to original images, ground truth, prediction results of CSG, and GCISG.

In contrast, our method can be applied to any general architecture with no additional preparation steps. In Fig. 6, we qualitatively compare with CSG [5] by visualizing the results of segmentation on validation images.

4.4 Object Detection

Datasets. We conduct object detection experiments on the source domain of Sim10k [22] to the target domain of Cityscapes. Sim10k dataset consists of 10,000 images with bounding box annotations on the car object. Images in Sim10k are rendered by the Grand Theft Auto V game engine, where the resolution of the images are 1914×1052. Since Cityscapes does not contain bounding box labels, we generate bounding box labels from polygon labels, following the method of [8].

Implementation details. We experiment on Faster R-CNN [37] as the base detector and ImageNet pre-trained ResNet-101 with FPN [26] as the backbone. We use an SGD optimizer with a learning rate of 0.001, and a batch size of 4 for 30 epochs. We use the same hyperparameter settings for the original Faster R-CNN except that we add color jittering for data augmentation. Since CSG [5] does not contain experiments on object detection, we implement the CSG framework for our experiments for a fair comparison.

Results. We evaluate the average precision of bounding boxes on each generalization method following the COCO [27] evaluation protocol. The results are shown in Table 7. Compared to the previous state-of-the-art syn-to-real generalization method [5], GCISG achieves around 2.1% points improvement in AP and 4.2% points improvement in AP50. In Fig. 7, we qualitatively compare with the CSG [5] by visualizing the results of the object detection on validation images.

4.5 Unsupervised Domain Adaptation

In this section, we demonstrate the effectiveness of GCISG for the unsupervised domain adaptation (UDA) task. We conduct experiments on class-balanced self-training (CBST) [51] framework, where we use the model trained with GCISG as a starting point for the adaptation.

We perform UDA experiments on the VisDA-17 image classification dataset. We follow the setting of [52], where we set the starting portion of the pseudo label p to 20%, and empirically add 5% to p for each epoch until it reaches 50%. Remark that, unlike previous syn-to-real generalization tasks, we freeze the normalization layers when training the source model, where we empirically found it better for the UDA task.

In Table 8, we present an accuracy on VisDA-17 validation dataset under various initialization methods for CBST framework. Compared to the baseline, our method remarkably boosts the performance by 7%, achieving 83.6%. Also, the model trained from GCISG outperforms that from ASG by 1%.

5 Conclusion

In this work, we present GCISG, which enhances the syn-to-real generalization by learning style-invariant representation and retaining the semantic knowledge of the ImageNet pre-trained model simultaneously. Through extensive experiments on VisDA-17 image classification, GTAV-Cityscapes semantic segmentation, and object detection, we demonstrate the effectiveness of our method.

Our work shows that we can extract useful information over the style changes that are useful for generalization. For future works, we aim to design style transformation methods that can better disentangle the style and content which suits our structural causal model. We believe that those style transformations can lead to better syn-to-real generalization under our method as they seek better causal invariance. We leave them for future work.

Acknowledgements. Kyungmin Lee is supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

1. Baker, N., Lu, H., Erlikhman, G., Kellman, P.J.: Deep convolutional networks do not classify based on global object shape. PLOS computational biology (2018)
2. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: ECCV (2018)
3. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
5. Chen, W., Yu, Z., Mello, S., Liu, S., Alvarez, J.M., Wang, Z., Anandkumar, A.: Contrastive syn-to-real generalization. In: ICLR (2021)
6. Chen, W., Yu, Z., Wang, Z., Anandkumar, A.: Automated synthetic-to-real generalization. In: ICML (2020)
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv:2003.04297 (2020)
8. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR (2018)
9. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: CVPR (2018)
10. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: CVPR (2021)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
12. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPR Workshops (2020)
13. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: CVPR (2019)
14. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: ECCV (2020)
15. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: Seed: Self-supervised distillation for visual representation. arXiv:2101.04731 (2021)
16. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016)
17. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 (2018)
18. Handa, A., Pătrăucean, V., Stent, S., Cipolla, R.: Scenenet: an annotated model generator for indoor scene understanding. In: ICRA (2016)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
21. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
22. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In: ICRA (2017)

23. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: ICML (2019)
24. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
25. Li, Z., Hoiem, D.: Learning without forgetting. In: TPAMI (2017)
26. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. arXiv:1612.03144 (2016)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, L.: Microsoft coco: Common objects in context. In: ECCV (2014)
28. Mahajan, D., Tople, S., Sharma, A.: Domain generalization using causal matching. In: ICML (2021)
29. Maximov, M., Galim, K., Leal-Taixe, L.: Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. In: CVPR (2020)
30. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet RGB-D: 5m photorealistic images of synthetic indoor trajectories with ground truth. arXiv:1612.05079 (2016)
31. Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., Blundell, C.: Representation learning via invariant causal mechanisms. arXiv:2010.07922 (2020)
32. Müller, M., Casser, V., Lahoud, J., Smith, N., Ghanem, B.: Sim4cv: A photorealistic simulator for computer vision applications. In: IJCV (2018)
33. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018)
34. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)
35. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv:1710.06924 (2017)
36. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: CVPR (2017)
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: TPAMI (2017)
38. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016)
39. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
41. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv:1910.10699 (2019)
42. Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: CVPR Workshops (2018)
43. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: ICCV (2019)
44. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: ICCV (2021)
45. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
46. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: ECCV (2020)

47. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A.L., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: ICCV (2019)
48. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:1612.03928 (2016)
49. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: ICML (2017)
50. Zheng, M., You, S., Wang, F., Qian, C., Zhang, C., Wang, X., Xu, C.: ReSSL: Relational self-supervised learning with weak augmentation. In: NIPS (2021)
51. Zou, Y., Yu, Z., Kumar, B.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018)
52. Zou, Y., Yu, Z., Liu, X., Kumar, B.V., Wang, J.: Confidence regularized self-training. In: ICCV (2019)