RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening

Sungha Choi*1,3 Sanghun Jung*2 Huiwon Yun1 Joanne T. Kim3 Seungryong Kim3 Jaegul Choo2

1LG AI Research 2KAIST 3Korea University 4Sogang University

Abstract

Enhancing the generalization capability of deep neural networks to unseen domains is crucial for safety-critical applications in the real world such as autonomous driving. To address this issue, this paper proposes a novel instance selective whitening loss to improve the robustness of the segmentation networks for unseen domains. Our approach disentangles the domain-specific style and domain-invariant content encoded in higher-order statistics (i.e., feature covariance) of the feature representations and selectively removes only the style information causing domain shift. As shown in Fig. 1, our method provides reasonable predictions for (a) low-illuminated, (b) rainy, and (c) unexpected scenes. These types of images are not included in the training dataset, where the baseline shows a significant performance drop, contrary to ours. Being simple yet effective, our approach improves the robustness of various backbone networks without additional computational cost. We conduct extensive experiments in urban-scene segmentation and show the superiority of our approach to existing work. Our code is available at this link1.

* indicates equal contribution
1 https://github.com/shachoi/RobustNet.

1. Introduction

When deploying deep neural networks (DNNs) trained on a given dataset (i.e., source domain) in real-world unseen data (i.e., target domain), DNNs often fail to perform properly due to the domain shift. Overcoming this issue is crucial, especially for safety-critical applications such as autonomous driving. In particular, real-world data consist of unexpected and unseen samples, for example, those images taken under diverse illumination, adverse weather conditions, or from different locations. It is generally impossible to model such a full data distribution with limited training data, so reducing the domain gap between source and target domains has been a long-standing problem in computer vision.

Domain adaptation (DA) is an approach to mitigate the performance degradation caused by such a domain gap [3, 12, 18, 11, 65, 39, 58, 41, 50]. Generally, DA focuses on adapting the source domain distribution to that of the target domain, but it requires access to the samples in the target domain, which limits their applicability. When we set the entire real world as a target domain, it is difficult in practice to obtain data samples that fully cover the target domain.

Domain generalization (DG) overcomes this limitation...
loss that alleviates the limitations of the existing whitening transformation for domain generalization, by selectively removing information that causes a domain shift while maintaining a discriminative power of feature within DNNs. Our method does not rely on an explicit closed-form whitening transformation, but implicitly encourage the networks to learn such a whitening transformation through the proposed loss function, thus requiring negligible computational cost. As illustrated in Fig. 2, our method selectively removes only those feature covariances that respond sensitively to photometric augmentation such as color transformation. Our experiments on urban-scene segmentation in DG settings, performed using several backbone networks, show evidence that our approach consistently boosts the DG performance.

The main contributions include the following:

- We propose an instance selective whitening loss for domain generalization, which disentangles domain-specific and domain-invariant properties from higher-order statistics of the feature representation and selectively suppresses domain-specific ones.
- Our proposed loss can easily be used in existing models and significantly improves the generalization ability with negligible computational cost.
- We apply the proposed loss to urban-scene segmentation in a DG setting and show the superiority of our approach over existing approaches in both a qualitative and quantitative manner.

2. Related Work

Domain adaptation and generalization
It is well known that significant labeling efforts are required so as to ensure the reliable performance of various tasks such as semantic segmentation [33, 1, 4, 64, 8]. To tackle this challenge, domain adaptation (DA) methods were proposed to transfer the knowledge learned from abundant labeled data (i.e., a source domain) to a target domain where labeled data are scarce. In contrast to DA, domain generalization (DG) methods assume that the model cannot access the target domain during training and aim to improve the generalization ability to perform well in an unseen target domain. Various approaches such as meta-learning [25, 2, 26, 31], adversarial training [27, 30, 44], autoencoder [15, 27], metric learning [10, 37], data augmentation [61, 16, 63] have been proposed to learn domain-agnostic feature representations. Recently, several studies [42, 53] have shown the effectiveness of exploiting both batch normalization (BN) [23] and instance normalization (IN) [57] within DNNs to solve the DG problem. These studies show that BN improves discriminative ability on features, while IN prevents overfitting on training data, so that generalization performance is improved on unseen domains by combining BN and IN. Especially, IBN-Net [42] shows a significant performance improvement with the marginal architectural modification...
that incorporates the IN layers through training on a single source domain, unlike most DG methods that require multiple source domains. This normalization based DG method is attractive because it can be applied as a complement to other DG methods based on multiple source domains.

**Semantic segmentation in DG** Based on the synthetic data such as GTAV [45] and SYNTHIA [47], numerous DA studies [41, 58, 51, 6, 65, 18, 56, 35, 62] have been proposed in semantic segmentation, but only a few DG studies [61, 42] address semantic segmentation, as the majority of the DG methods mainly focused on image classification. DA, which can access the target domains, generally has better performance than DG, but DG methods that can handle an arbitrary unseen domain without access to the target domain are mandatory in the real world. This paper focuses on the DG method practically helpful in semantic segmentation where various conditions exist such as adverse weather, diverse illumination, location differences, and so on.

**Feature covariance** The seminal studies [13, 14] have demonstrated that feature correlations (i.e., a gram matrix or covariance matrix) take style information of images. Since then, numerous studies exploit the feature correlation in style transfer [28], image-to-image translation [7], domain adaptation [49, 55] and networks architecture [34, 43, 20, 54]. Especially, the whitening transformation that removes feature correlation and makes each feature have unit variance, has been known to help to remove the style information from the feature representations [28, 43, 7]. Our work explores the whitening transformation to improve domain generalization performance. To the best of our knowledge, this is the first attempt to apply whitening to DG.

### 3. Preliminaries

**Whitening transformation (WT)** Let $X \in \mathbb{R}^{C \times HW}$ denote the intermediate feature map, where $C$ is the number of channels, $H$ and $W$ are the spatial dimensions of the feature map, height and width, respectively. WT is a linear transformation that makes the variance term of each channel equal to one and the covariances between each pair of channels equal to zero. A whitening-transformed feature map $\tilde{X}$ from $X$ satisfies that $\tilde{X} \cdot \tilde{X}^\top = (HW) \cdot I \in \mathbb{R}^{C \times C}$, where $I$ denotes the identity matrix, and can be computed as

$$\tilde{X} = \Sigma^{-\frac{1}{2}}_\mu (X - \mu \cdot 1^\top), \quad (1)$$

where $1 \in \mathbb{R}^{HW}$ is a column vector of ones, and $\mu$ and $\Sigma_\mu$ are the mean vector and the covariance matrix, respectively, i.e.,

$$\mu = \frac{1}{HW} X \cdot 1 \in \mathbb{R}^{C \times 1}, \quad (2)$$

$$\Sigma_\mu = \frac{1}{HW} (X - \mu \cdot 1^\top) (X - \mu \cdot 1^\top)^\top \in \mathbb{R}^{C \times C}. \quad (3)$$

Since the covariance matrix $\Sigma_\mu$ can be further eigen-decomposed such that $Q \Lambda Q^\top$, where $Q \in \mathbb{R}^{C \times C}$ is the orthogonal matrix of eigenvectors, and $\Lambda \in \mathbb{R}^{C \times C}$ is the diagonal matrix that contains each eigenvalue of the corresponding eigenvector from $Q$, we can calculate an inverse square root of the covariance matrix $\Sigma^{-\frac{1}{2}}_\mu$ as

$$\Sigma^{-\frac{1}{2}}_\mu = Q \Lambda^{-\frac{1}{2}} Q^\top. \quad (4)$$

It has been known that WT can effectively remove style information by being applied to each instance in style transfer [28].

**Limitations of WT** We can compute the whitening transformation matrix $\Sigma^{-\frac{1}{2}}_\mu$ analytically through Eq. (4), but eigenvalue decomposition is computationally expensive, leading to slow training and inference speed and prevents the gradient back-propagation [20, 7]. To alleviate these problems, previous studies have shown that the goal of WT can be achieved without the eigen-decomposition through the whitening loss [7] or approximating the whitening transformation matrix using Newton’s iteration [21, 20, 43].

Especially, GDWCT [7] proposes the deep whitening transformation (DWT) that implicitly makes the covariance matrix $\Sigma_\mu$ close to the identity matrix $I$ by means of the loss defined as

$$L_{DWT} = \mathbb{E}[\|\Sigma_\mu - I\|_1], \quad (5)$$

where $\mathbb{E}$ denotes the arithmetic mean. GDWCT applies this loss to image-to-image translation for more significant style changes than other methods [22, 24] of aligning only the first-order statistics (i.e., channel-wise mean and variance). However, applying these alternative methods of WT to DG is not straightforward. Whitening all covariance elements may diminish feature discrimination [43, 59] and distort the boundary of an object [29, 28] because domain-specific style and domain-invariant content are simultaneously encoded in the covariance of the feature map.

### 4. Proposed Method

This section presents our approach to solve the domain generalization problem through whitening the feature representation by mitigating undesirable effects of a whitening transformation. Our method disentangles the covariance into the encoded style and content so that only the style information can be selectively removed, thus increasing the domain generalization ability. We firstly propose an instance whitening and instance-relaxed loss in Section 4.1 and then finally propose our novel instance selective whitening loss in Section 4.3.

#### 4.1. Instance Whitening Loss

This subsection describes a series of steps to transform the input feature into the whitening transformed feature as shown in Fig. 3. Note that our method is applied to each instance, not to a mini-batch. Let $\Sigma_{\mu(i, i)}$ denote a diagonal
Main networks

Figure 3. Overall process of our proposed method. (a) Instance standardization. (b) Deriving a covariance matrix from a standardized feature map. (c) Leaving only the covariance to which the whitening loss is applied. (d) Applying the criterion that measures the mean absolute error between the remaining covariance values and zero. No additional computation is required for inference as the operations in red are used only for training. Notations: X: intermediate feature map, Xs: standardized feature map, Σs: covariance matrix of the standardized feature map, M: matrix for masking, LIW: our proposed instance whitening loss.

Element (i, i) and Σμ(i, i) denote an off-diagonal element (i, j) of the covariance matrix Σs of the intermediate feature map, where 0 ≤ i, j < C, i ≠ j. The DWT loss in Eq. (5) can be decomposed as

||Σs(i, i)−1||1 ||xι||HW−1||1 = ||xι||HW−1||1 ||xι||HW−1||1, (6)

||Σs(i, j)||1 ||xι||HW−1||1 = ||xι||HW−1||1 ||xι||HW−1||1, (7)

where xι ∈ RHW denotes the i-th channel of the intermediate feature map X ∈ RCHxHW. Note that Eq. (6) applies to the diagonal elements, and Eq. (7) applies to the off-diagonal elements of the covariance matrix. The optimization process for the whitening loss should minimize both Eq. (6) and Eq. (7) simultaneously, but there exists a limitation on it. The scale of each channel (i.e., ||xι||) is forced to increase to the value of √HW by Eq. (6) and decrease to zero by Eq. (7). Therefore, forcing the diagonal and off-diagonal of the covariance matrix to be one and zero, respectively, conflicts with each other, so it is difficult to optimize both at the same time.

To address this issue, the feature map X can first be standardized into Xs through an instance normalization [57]:

Xs = diag(Sμ)−1/2 (X − μ · 1T), (8)

where ⋄ is an element-wise multiplication, and diag(Sμ) ∈ RCHx1 denotes the column vector consisting of diagonal elements in the covariance matrix. Note that each diagonal element is copied along with the spatial dimension HW for element-wise multiplication. Since the scale of each feature vector is already fixed as the unit value after the instance standardization, the whitening loss only affects the cos θ term in Eq. (7). In the end, this approach fits the purpose of the whitening transformation to decorrelate the features.

After standardization of the intermediate feature map, the covariance matrix is calculated as

Σs = 1/MTr(Xs)XsT ∈ RCHxC, (9)

where Xs is the standardized feature map. Thanks to the standardization process, diagonal elements of the covariance matrix are already set as unit values. Thus, we only need to make the off-diagonals of the covariance matrix close to zero, which makes it easy to optimize for the whitening process, and the aforementioned conflict can thus be resolved. Since the covariance matrix is symmetric, the loss can be applied only to the strict upper triangular part.

Our instance whitening (IW) loss is formulated as

LIW = E[∥Σs ◦ M∥1], (10)

where E denotes the arithmetic mean and M ∈ RCHxCH denotes a strict upper triangular matrix, i.e.,

Mij = {0, if i ≥ j
1, otherwise 0 ≤ i, j < C. (11)

4.2. Margin-based relaxation of whitening loss

The instance whitening loss (Eq. (10)) suppresses all covariance elements to zero, so it can adversely affect the discriminative power of features within DNNs. To address this issue, we propose an instance-relaxed whitening (IRW) loss to sustain the covariance elements essential in maintaining the discriminative power. The IRW loss is designed so that the expected value of the total covariance lies within a specified margin δ rather than being close to zero, i.e.,

LIRW = max(E[∥Σs ◦ M∥1] − δ, 0) (12)

The loss LIRW allows the covariance to have a certain level of values, so it gives room to keep discriminative features intact. The empirical effect of the IRW loss can be found in Section 5.2.1. It shows better performance compared to the IW loss not including margin δ (Eq. (10)). Nonetheless, it may not be sufficient because we cannot guarantee that only the covariance useful for generalization performance remains through the margin relaxation.

4.3. Separating Covariance Elements

To further improve our approach, we need to separate the covariance terms into two groups: domain-specific style and domain-invariant content. We propose to selectively suppress only the style-encoded covariances that cause the
domain shift. Assuming that the domain shift includes
changes in color and blurriness, we simulate the domain shift through photometric augmentation such as color jittering and Gaussian blurring.

First, we add only the instance standardization layer into
the networks (Fig. 3(a)) and train them during the $n$
initial epochs without the whitening loss to get the pure statistics of the covariance matrices from training images. $n$ is
a hyper-parameter, which we empirically set to 5. Afterwards, we extract two covariance matrices by inferring from
two input images, namely an original and a photometric-
transformed image, and calculate the variance matrix from the differences between two different covariance matrices. Formally, the variance matrix $\mathbf{V} \in \mathbb{R}^{C \times C}$ is defined as

$$\mathbf{V} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2,$$

(13)

from mean $\mu_{\Sigma_i}$ and variance $\sigma_i^2$ for each element from two different covariance matrices of the $i$-th image, i.e.,

$$\mu_{\Sigma_i} = \frac{1}{2} (\Sigma_s(x_i) + \Sigma_s(\tau(x_i)))$$

(14)

$$\sigma_i^2 = \frac{1}{2} \left( (\Sigma_s(x_i) - \mu_{\Sigma_i})^2 + (\Sigma_s(\tau(x_i)) - \mu_{\Sigma_i})^2 \right)$$

(15)

where $N$ denotes the number of image samples, $x_i$ is the $i$-th image sample, $\tau$ is a photometric transformation, and $\Sigma_s(\cdot)$ extracts the covariance matrix of the intermediate feature map from an input image. As a result, $\mathbf{V}$ consists of elements of the variance of each covariance element across various photometric transformations.

We assume that the variance matrix $\mathbf{V}$ implies the sensitivity of the corresponding covariance to the photometric transformation. This means that the covariance elements with high variance value contain the domain-specific style such as color and blurriness. To identify such elements, we apply $k$-means clustering on the strict upper triangular elements $V_{i,j}$ ($i < j$) of the variance matrix $\mathbf{V}$ to assign the elements into $k$ clusters $C = \{c_1, c_2, \ldots, c_k\}$ with respect to the value. Next, we split the $k$ clusters into two groups, $G_{low} = \{c_1, \ldots, c_m\}$ with low variance value and $G_{high} = \{c_{m+1}, \ldots, c_k\}$ with high variance value. The hyper-parameters $k$ and $m$ are empirically set to 3 and 1, respectively. More details can be found in the supplementary Section A.2. We assume that $G_{high}$ contains the domain-specific style and $G_{low}$ contains domain-invariant content.

Finally, we propose an instance selective whitening (ISW) loss that selectively suppresses only to the style-encoded covariances. Let the mask matrix $\mathbf{M}$ in Eq. (11) change to $\tilde{\mathbf{M}} \in \mathbb{R}^{C \times C}$ for the ISW loss as

$$\tilde{M}_{i,j} = \begin{cases} 1, & \text{if } V_{i,j} \in G_{high} \\ 0, & \text{otherwise} \end{cases}$$

(16)

The ISW loss is defined as

$$L_{ISW} = \mathbb{E}[||\Sigma_s \odot \tilde{\mathbf{M}}||_1].$$

(17)

The networks continue training for the remaining epochs incorporating the proposed ISW loss.

### 4.4. Network architecture with proposed ISW loss

IBN-Net [42] has explored a number of ResNet [17]-based architectures to combine instance normalization with batch normalization and proposed several IBN blocks based on a residual block (Fig. 5(a)). Among the proposed blocks, IBN-b, which adds an instance normalization layer right after the addition operation of a residual block (Fig. 5(b)),

![Figure 5. Architecture comparison with other methods: (a) the original residual block [17]; (b) IBN-b [42] combining instance normalization with batch normalization; (c) IterNorm [21] employing Newton’s iterations for efficient whitening; (d) Our proposed ISW loss applied to instance normalization.](image)
shows the best generalization performance on semantic segmentation tasks. After all, they add three instance normalization layers after the first three convolution groups (i.e., conv1, conv2, and conv3). We follow this architectural approach as our baseline. As shown in Fig. 5(d), we simply add our proposed ISW loss to the instance normalization layer. Our loss in total is described as

\[ \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \left( \frac{1}{L} \sum_{i} \mathcal{L}_{i}^{\text{ISW}} \right), \]  

where \( \lambda \) denotes the weight of our ISW loss and is empirically set to 0.6, \( \mathcal{L}_{\text{task}} \) is the task loss (e.g., a per-pixel cross-entropy loss for semantic segmentation), \( i \) indicates the layer index, and \( L \) is the number of layers to which the ISW loss is applied. The hyper-parameter \( \lambda \) is analyzed in the supplementary Section A.2. \( L \) is set to three by following IBN-Net. An affine transformation is not used since the subsequent convolution operation after a whitening transformation can do the equivalent job, and empirically, we found no performance gain by explicitly adding the affine transformation.

5. Experiments

This section describes the experimental setup and presents evaluation results to assess the effectiveness of our proposed methods on semantic segmentation with comparison to other methods. Furthermore, we provide an in-depth analysis of our results including the covariance matrices.

5.1. Experimental Setup

We train our model on several datasets (e.g., Cityscapes) and show its performance on other datasets (e.g., BDD-100K, Mapillary, GTAV, and SYNTHIA) to measure the generalization capability on unseen domains. For fair comparisons with other normalization techniques, we re-implement IBN-Net [42] and IterNorm [21] on our baseline models and compare them with our methods. As described in Section 4.4, our proposed loss can easily be added to existing models, so we apply our methods to various backbone networks such as ResNet [17], ShuffleNetV2 [36] and MobileNetV2 [52] and show wide applicability of the proposed methods. For all the quantitative experiments, mean Intersection over Union (mIoU) is used to measure the segmentation performance.

5.1.1 Implementation details

We adopt DeepLabV3+ [5] for a semantic segmentation architecture, and SGD optimizer with an initial learning rate of 1e-2 and momentum of 0.9 is used. Besides, we follow the polynomial learning rate scheduling [32] with the power of 0.9. We train all the models for 40K iteration, except for multi-source models, which are trained for 110K iterations. To prevent the model from overfitting, color and positional augmentations such as color jittering, Gaussian blur, random cropping, random horizontal flipping, and random scaling with the range of [0.5, 2.0] are conducted. For the photometric transformation in ISW, we apply color jittering and Gaussian blur. Also, as suggested by IBN-Net, we add three instance normalization layers after the first three convolution groups and apply our proposed loss. Further details are provided in the supplementary Section A.3.

5.1.2 Datasets

To verify the generalization capability of our methods, we conduct the experiments on five different datasets.

Real-world datasets Cityscapes [9] is a large-scale dataset containing high-resolution (e.g., 2048 × 1024) urban scene images collected from 50 different cities in primarily Germany. It provides 3,450 finely-annotated images and 20,000 coarsely-annotated images. We use only a finely-annotated set for training and validation. BDD-100K [60] is another real-world dataset that contains diverse urban driving scene images with the resolution of 1280 × 720. The images are collected from various locations in the US. For a semantic segmentation task, 7,000 training and 1,000 validation images are provided. The last real-world dataset we use is Mapillary [40], a diverse street-view dataset consisting of 25,000 high-resolution images with a minimum resolution of 1920 × 1080 collected from all around the world.

Synthetic datasets GTAV [45] is a large-scale dataset containing 24,966 driving-scene images generated from Grand Theft Auto V game engine. It has 12,403, 6,382, and 6,181 images of size 1914 × 1052 for a train, a validation, and a test set, respectively. It has 19 object categories compatible with Cityscapes. Also, we use SYNTHIA [48], composed of photo-realistic synthetic images containing 9,400 samples with a resolution of 960 × 720.

5.2. Quantitative Evaluation

This subsection provides ablation studies, the comparisons of our results against other normalization methods, the evaluation on multiple source domains, and the analysis of computational cost. Since the experiments follow domain generalization settings, the model cannot access any datasets other than the source data.

5.2.1 Effectiveness of instance selective whitening loss

To verify the effectiveness of our methods, we conduct comparisons with other normalization methods and ablation studies on instance whitening (IW), instance-relaxed whitening (IRW), and instance selective whitening (ISW).

Note that all the experiments in this subsection are performed three times and averaged for fair comparisons.

Table 1 shows the generalization performance of the models trained on GTAV dataset. ISW outperforms other methods on all datasets except the source dataset (i.e., GTAV). Especially, ISW shows a significant improvement on real-world datasets (i.e., Cityscapes, BDD-100K, and
shows our model trained on multiple synthetic domains. Table 4 shows our model trained on multiple datasets performs better than other models due to its generalization ability by extracting domain-invariant features during training.

Table 2. Comparison of mIoU(%). The models are trained on GTA V train set. The backbone networks of the first group are ShuffleNetV2 \[36\] and the second group is MobileNetV2 \[52\].

<table>
<thead>
<tr>
<th>Models (Cityscapes)</th>
<th>B</th>
<th>M</th>
<th>G</th>
<th>S</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>44.96</td>
<td>51.68</td>
<td>42.55</td>
<td>23.29</td>
<td>77.51</td>
</tr>
<tr>
<td>(^{†})SW [43]</td>
<td>48.49</td>
<td>55.82</td>
<td>44.87</td>
<td>26.10</td>
<td>77.30</td>
</tr>
<tr>
<td>(^{†})IBN-Net [42]</td>
<td>48.56</td>
<td>57.04</td>
<td>45.06</td>
<td>26.14</td>
<td>76.55</td>
</tr>
<tr>
<td>(^{†})IterNorm [21]</td>
<td>49.23</td>
<td>56.26</td>
<td>45.73</td>
<td>25.98</td>
<td>76.02</td>
</tr>
<tr>
<td>Ours (IW)</td>
<td>48.19</td>
<td>58.90</td>
<td>45.21</td>
<td>25.81</td>
<td>76.06</td>
</tr>
<tr>
<td>Ours (IRW)</td>
<td>48.67</td>
<td>59.20</td>
<td>45.64</td>
<td>26.05</td>
<td>76.13</td>
</tr>
<tr>
<td>Ours (ISW)</td>
<td>50.73</td>
<td>58.64</td>
<td>45.00</td>
<td>26.20</td>
<td>76.41</td>
</tr>
</tbody>
</table>

Table 3. Comparison of mIoU(%). The models are trained on GTA V train set. The backbone networks of the first group are ShuffleNetV2 \[36\] and the second group is MobileNetV2 \[52\].

<table>
<thead>
<tr>
<th>Models (GTAV)</th>
<th>C</th>
<th>B</th>
<th>M</th>
<th>G</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>22.20</td>
<td>29.60</td>
<td>7.40 (^{†})</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>(^{†})IBN-Net [42]</td>
<td>32.45</td>
<td>37.42</td>
<td>4.97 (^{†})</td>
<td>26.73</td>
<td>32.14</td>
</tr>
<tr>
<td>Ours (ISW)</td>
<td>28.95</td>
<td>36.58</td>
<td>7.63 (^{†})</td>
<td>25.14</td>
<td>35.20</td>
</tr>
</tbody>
</table>

Table 4. Comparison of mIoU(%). The models are trained on multiple synthetic domains. The backbone is ResNet-50 with an output stride of 16. \(^{†}\) denotes re-implemented models.

5.2.2 Comparison with other DG and DA methods

This subsection compares our method with two existing DG methods on semantic segmentation task, based on the results reported in the papers \[42, 61\]. DRPC \[61\] proposes a domain randomization method, which maps the synthetic images to multiple auxiliary real domains using image-to-image translation with the style of real images (e.g., ImageNet). As shown in Table 5, our model gains the largest performance increase on average, compared to other methods such as IBN-Net \[42\] and DRPC \[61\]. Our method shows a large amount of performance improvement on BDD-100K and Mapillary datasets that involve significantly more diverse driving scenes than Cityscapes.

In addition, we compare the result of our method with those reported from several domain adaptation methods. See the supplementary Section A.1.
### 5.2.3 Computational cost analysis

To ensure our method requires no additional computational cost, we report the number of parameters, GFLOPS, and inference time. As seen in Fig. 5, all the models in Table 6 share the same network architecture, but with different normalization methods. As shown in Table 6, our approach performs a whitening transformation without additional computational cost.

### 5.3. Qualitative Analysis

#### Comparison of covariance matrices

To show how the covariance matrix is selectively whitened, we visualize the covariance matrix of intermediate feature maps from IBN-Net [42] and our model with ISW. As shown in Fig. 6, the first pair of covariance matrices are from the first convolution layer and the others are from the second convolution layer. Note that the style information mainly exists in the early layers of the network as pointed out in IBN-Net. Moreover, the style information is encoded as a form of the features covariance as revealed in previous studies [13, 14]. Hence, the covariance matrices are sparser at the second pair, compared to the first ones. By comparing the covariance maps from IBN-Net and ISW, we can find the ones from ours are whitened but a small number of covariance elements remain large, showing our ISW selectively eliminates the covariance.

#### Reconstructing images with whitened features

For in-depth analysis, we reconstruct input images from the whitened feature maps of our ISW model. For the experiment, we adopt U-Net [46] as reconstruction networks. To newly train a decoder, we append the decoder to the backbone of a pre-trained baseline and train the decoder. We then replace the backbone network with the pre-trained ISW model. As seen in Fig. 7, generated images preserve the relevant content information for segmentation while the style information such as illumination and colors is suppressed. These examples support the validity of our approach that selectively suppresses the style information.

### 6. Discussions

In this section, we discuss potential issues and improvements of our approach for further research.

**Affine parameters.** Most of the normalization layers contain affine parameters to recover the original distribution and enhance the representation of a network. We attempted to deploy this by adding affine parameters or a $1 \times 1$ convolution layer after the normalization layer incorporating our proposed whitening loss. Despite our effort, this approach did not improve our method. We conjecture it is because affine parameters or a $1 \times 1$ convolution layer do not have sufficient complexity in recovering the original distribution.

**Photometric transformation.** Our method adopted photometric transformation to separate the style and content information, where we found that applying color transform and Gaussian blur does not harm the content information. We expect our approach can be further improved by exploring various photometric augmentation techniques.

### 7. Conclusions

This paper proposed a novel instance selective whitening (ISW) loss, which facilitates disentangling the covariances of the intermediate features into the style- and content-related ones and suppressing only the former to learn the domain-invariant feature representation. We focused on solving the domain generalization problem in urban-scene segmentation, which has practical impact when deployed in the wild but has not been studied much. In this regard, we strive to promote the importance of the domain generalization and inspire new research paths in this area.

**Acknowledgments** This work was partially 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) and No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2019R1A2C4070420).

---

<table>
<thead>
<tr>
<th>Models</th>
<th># of Params</th>
<th>GFLOPS</th>
<th>Inference Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>45.082M</td>
<td>554.31</td>
<td>10.48</td>
</tr>
<tr>
<td>IBN-Net [42]</td>
<td>45.083M</td>
<td>554.31</td>
<td>10.51</td>
</tr>
<tr>
<td>IterNorm [21]</td>
<td>45.081M</td>
<td>554.31</td>
<td>40.31</td>
</tr>
<tr>
<td>Ours</td>
<td>45.081M</td>
<td>554.31</td>
<td>10.43</td>
</tr>
</tbody>
</table>

Table 6. Comparison of computational cost. Tested with the image size of $2048 \times 1024$ on NVIDIA A100 GPU. The inference time is averaged over 500 trials. † denotes re-implemented models.

---

Figure 6. Visualization of covariance matrices extracted from IBN-Net and our model. The first and the second pairs are extracted from the first and the second convolution layers, respectively.

Figure 7. Reconstructed images from ISW-whitened feature maps using U-Net; the first row: a baseline backbone, the second row: an ISW model backbone. The image contents are properly maintained while the style such as illumination and colors vanish.
References


[20] Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3


[28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature


[56] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schultler, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker.


