An Instance Normalization Transformer for Generalized Driving-Scene Segmentation

Qi Bi   Shaodi You   Theo Gevers

Computer Vision Research Group, University of Amsterdam

Amsterdam, the Netherlands

{q.bi, s.you, th.gevers}@uva.nl

Abstract

Generalizability is highly demanded for semantic segmentation, especially for real world applications such as autonomous driving. Although Vision Transformers (ViT) have shown their potential in different computer vision applications compared to CNN-based methods, they are rarely used in the domain of generalized segmentation.

Therefore, in this paper, we propose a novel instance normalization Transformer (INFormer). To the best of our knowledge, instance normalization has not been explored so far using patch-wise ViT embeddings. To this end, we propose a progressive normalization strategy, which applies normalization in both the encoding and decoding stages. After feature encoding, the image representation is directly implemented using instance normalization. During the decoding stage, the image features for each scale are normalized and fused into the Transformer decoder in a progressive manner. Large-scale experiments, considering a variety of driving-scene scenarios, show that the proposed INFormer significantly outperforms existing CNN based domain generalized semantic segmentation methods by up to 12.79% mIoU.

1. Introduction

Semantic segmentation in driving scenarios is particularly challenging because the environment may change drastically including changes in weather and lighting conditions, and variations in landscapes [2, 6, 20, 32]. Existing segmentation models are usually trained on well-illuminated datasets and are therefore they may not be robust in other domains. Domain generalized semantic segmentation [4, 11, 22] benchmarks are thus proposed to systematically study this problem.

Vision Transformers (ViT's) show stronger feature generalization capabilities than CNN's [7, 16, 17]. Recently, ViT is applied successfully in semantic segmentation including typical methods such as SegFormer [36] and Mask2Former [3]. However, existing domain generalized semantic segmentation approaches heavily rely on convolutional neural networks (CNN's) [4, 11, 22–24, 24, 37]. So far ViT has not be explored in the domain of generalized semantic segmentation.

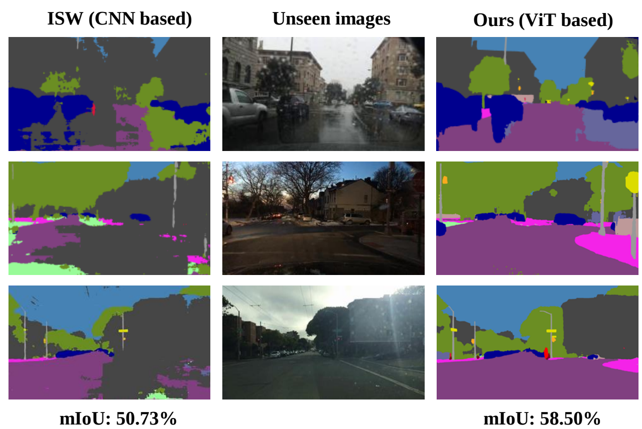

Figure 1. The proposed **I**nstance **N**ormalization Trans**Former** (INFormer) is a VIT based domain generalized segmentation method for driving scenes. It shows a significant performance gain compared to CNN based domain generalized segmentation methods, *e.g.*, ISW [4].

While **I**nstance **N**ormalization (IN) is well researched using CNN's, it has not been explored by ViT's [11, 22, 24]. Therefore, in this paper, a novel **I**nstance **N**ormalization Trans**Former** (INFormer) is proposed for domain generalized semantic segmentation. Firstly, the feature embedding from the transformer encoder is implemented with the IN transformation. It allows the high-level feature embedding to be more robust to the style variation. Then, in the decoding stage, the image features of each scale are implemented by the instance normalization, and progressively fused into the Transformer decoder. In this way, the IN keeps playing its role during the up-sampling process, so that more gener-

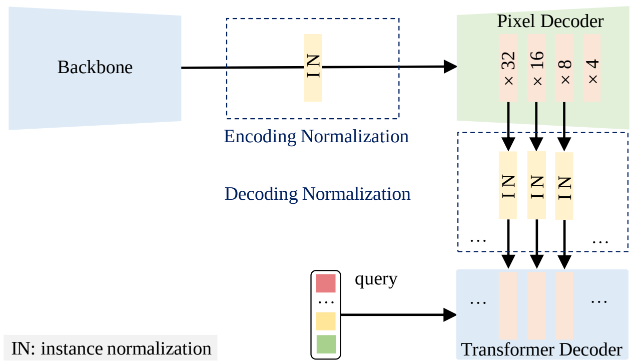

Figure 2. An overview of the training pipeline for the proposed IN Transformer (INFormer). The key idea is that the IN is used in both the encoding and decoding stage, and the fusion in the decoding stage is in a progressive manner.

alized high resolution dense predictions are made.

Large-scale experiments for different domain generalized segmentation scenarios show that the proposed INFormer outperforms state-of-the-art CNN based methods by a large margin i.e. 12.79% mIoU improvement. In addition, ablation studies show the necessity of implementing IN transformation repeatedly in the encoding and decoding stages. The visualized prediction in the target domains demonstrates the reliability of the proposed INFormer for domain generalized segmentation, compared to existing state-of-the-art CNN based methods.

Our contribution is summarized as follows.

- To the best of our knowledge, this is the first approach using ViT for domain generalized semantic segmentation.
- We propose a **I**nstance **N**ormalization Trans**Former** (INFormer) for the domain generalization semantic segmentation task.
- Extensive experiments show that the proposed INFormer leads to a 12.79% mIoU improvement compared to existing CNN based domain generalized semantic segmentation methods.

2. Related Work

Domain Generalization Extensive efforts of domain generalization under no task-specific scenarios are made in both machine learning and computer vision community. Specifically, Zhou *et al*. [44] provide an extensive summary of domain generalization on a variety of vision tasks. Dou *et al*. [8] introduce a model-agnostic learning scheme to preserve domain generalized semantic features. Harary *et al*. [9] consider the domain generalization in an unsupervised manner by learning a *domain bridge*. Hu *et al*. [10] propose a domain generalization framework for image retrieval in an unsupervised setting. Zhou *et al*. [45] propose a framework to generalize to new homogeneous domains. Xu *et al*. introduces a domain generalization method based on a Fourier-based augmentation strategy and a dual-formed consistency loss. Qiao *et al*. [29] and Peng *et al*. [26] investigate how to learn domain generalization from a single source domain.

Meanwhile, methods such as entropy regularization [41], common-specific low-rank decomposition [27], casual matching [18], extrinsic-intrinsic interaction [35], balance invariance [1], batch normalization embeddings [33] and multiple latent domain modeling [19] are proposed.

Domain Generalized Semantic Segmentation Domain generalized semantic segmentation can be regarded as a boarder extension of the prior unsupervised domain adaption segmentation task [24, 25, 39], but demands more generalization ability of a model on a variety of target domains.

Despite some efforts on leveraging in-the-wild images [28], scribble images [34] and multi-source images [13, 14] for domain generalized segmentation, most attention in the vision community is still in generalized segmentation under driving-scenes [5, 21, 30, 31, 38].

Generally, domain generalization segmentation models use either normalization transformation (*e.g*., IBN [22], instance normalization [11], SAN [24]) or whitening transformation (*e.g*., IW [23], ISW [4], DIRL [37], SAW [24]) on the training domain, so that the model can better generalize on the target domains. Other more advanced domain generalization segmentation methods usually leverage external images for more diverse styles [15, 42, 43], or leverage the content consistency on multi-scale features [40].

3. Methodology

3.1. Encoding Normalization

Assume that the image feature from a Transformer encoder is denoted as $\mathbf{X} \in \mathbb{R}^{(W \cdot H) \times C}$, where C is the number of channels. Along the channel-wise, $\mathbf{X}$ can be divided by $\mathbf{X} = [\mathbf{X}^{(W \cdot H),1}, \cdots, \mathbf{X}^{(W \cdot H),C}]$.

The instance normalization is computed on $\mathbf{X}$ so that the feature representation of each individual channel is normalized. This process is defined by

$$\mathbf{X}'^{(W \cdot H),c} = \frac{\mathbf{X}^{(W \cdot H),c} - \mu}{\sigma + \epsilon} \cdot \gamma + \beta, \tag{1}$$

$$\mu = \frac{1}{C} \sum_{c=1}^{C} \mathbf{X}^{(W \cdot H),c}, \sigma = \sqrt{\frac{1}{C} \sum_{i=1}^{C} (\mathbf{X}^{(W \cdot H),c} - \mu)^2}, \tag{2}$$

where $c = 1, 2, \cdots, C$.

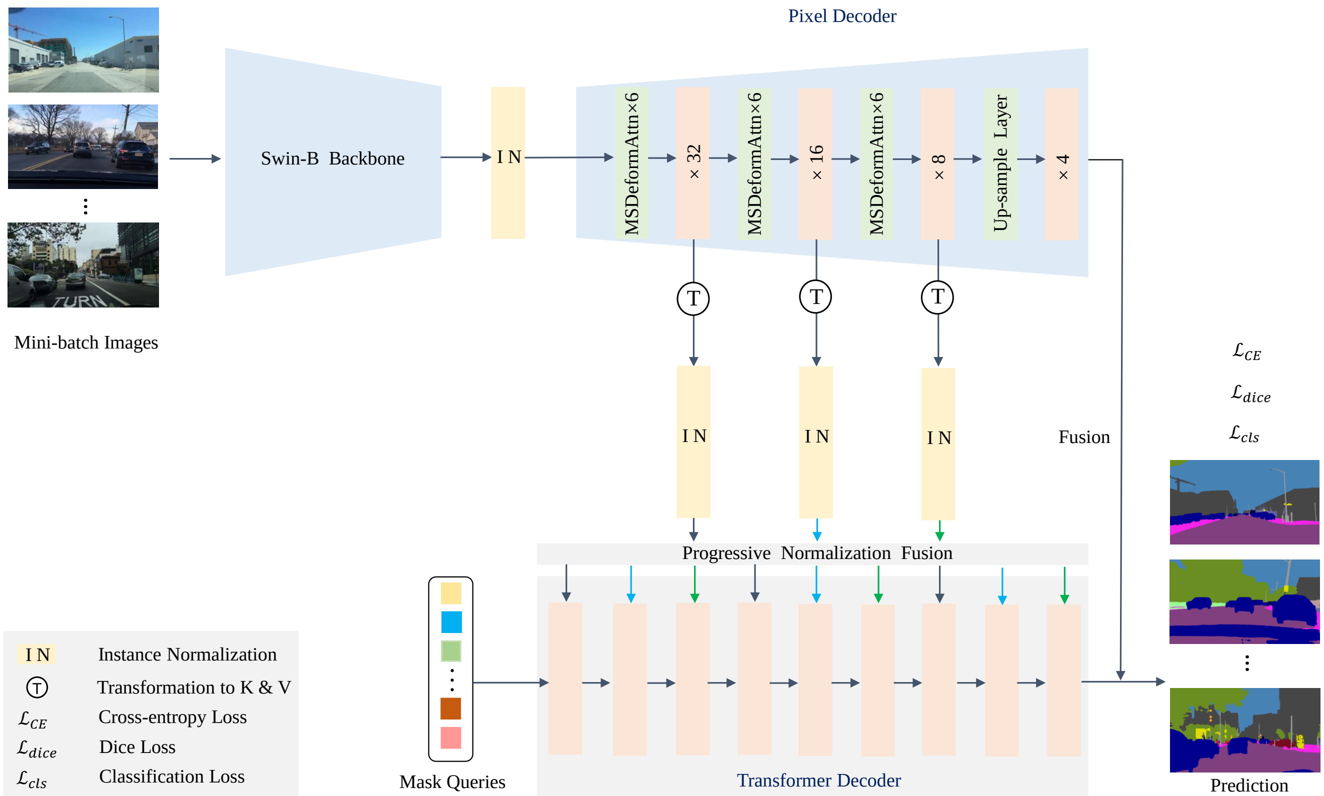

Figure 3. Technical framework of the proposed **I**nstance **N**ormalization Trans**Former** (INFormer) for domain generalized semantic segmentation. The Swin-B backbone and the Multi-scale deformable attention decoder are directly inherited from the Mask2Former segmentation backbone [3]. To learn generalized features, the features after encoding and decoding are both computed by the *instance normalization*. In the Transformer decoding stage, all image features for each scale are normalized, and are progressive fused into the Transformer decoder (in black, blue and green arrow.)

The features after normalization are fed into the Transformer decoder for subsequent processing, denoted by $\mathbf{X}^{'} = [\mathbf{X}^{'(W\cdot H),1}, \cdots, \mathbf{X}^{'(W\cdot H),C}]$.

3.2. Decoding Normalization

Modern vision Transformers (ViT) focus on the self-attention mechanism for stronger feature representations. For the segmentation task, the common paradigm (*e.g.*, SegFormer [36], Mask2Former [3]) is to extract features for dense prediction from a set of masks. To learn masks preserving more generalized features, the normalized features are used as input.

Let $\mathbf{X}_l \in \mathbb{R}^{N\times C}$ denote the features of the l^{th} layer in a Transformer decoder, where N is the number of semantic categories. Then, a standard masked self-attention mechanism for segmentation is computed as

$$\mathbf{X}_l = \mathrm{softmax}(\mathcal{M}_{l-1} + \mathbf{Q}_l\mathbf{K}_l^{\mathsf{T}})\mathbf{V}_l + \mathbf{X}_{l-1}, \tag{3}$$

where $\mathcal{M}_{l-1}$ is a binary mask to filter the foreground regions of an image, as detailed in [3]. Also, $\mathbf{Q}_l \in \mathbb{R}^{N\times C}$ denote the query features transformed from $\mathbf{X}_{l-1}$. $\mathbf{K}_l, \mathbf{V}_l \in \mathbb{R}^{(W\cdot H)\times C}$ denotes the key and value for $\mathbf{X}_{l-1}$. They are both computed for a certain image feature from the pixel decoder. For simplicity and clarity, in this subsection, the image feature (for $\mathbf{K}_l, \mathbf{V}_l$) before and after instance normalization is denoted by $\mathbf{F}_l$ and $\mathbf{F}'_l$, respectively. This process is computed as

$$\mathbf{F}_l^{'(W\cdot H),c} = \frac{\mathbf{F}_l^{(W\cdot H),c} - \mu_F}{\sigma_F + \epsilon} \cdot \gamma + \beta, \tag{4}$$

$$\mu_F = \frac{1}{C}\sum_{c=1}^{C}\mathbf{F}_l^{(W\cdot H),c}, \sigma_F = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(\mathbf{F}_l^{(W\cdot H),c} - \mu_F)^2}. \tag{5}$$

Then, based on $\mathbf{F}'_l$, two linear layers f_l^k and f_l^v are used to computed the key and value. It allows the key and value to carry more normalized image features. Let $\mathbf{K}'_l$ and $\mathbf{V}'_l$ denote the key and value from the normalized image features. This computation is given by

$$\mathbf{K}'_l = f_l^k(\mathbf{F}'_l), \tag{6}$$

$$\mathbf{V}'_l = f_l^v(\mathbf{F}'_l). \tag{7}$$

Then, in the proposed ViT for domain generalized segmentation, the masked self-attention mechanism is defined by

$$\mathbf{X}_l = \mathrm{softmax}(\mathcal{M}_{l-1} + \mathbf{Q}_l \mathbf{K}_l^{'\mathsf{T}})\mathbf{V}_l^{'} + \mathbf{X}_{l-1}, \quad (8)$$

where $\mathbf{X}_0$ is the output from the Transformer encoder, which is denoted by $\mathbf{X}^{'}$ in Sec. 3.1.

3.3. Progressive Normalized Fusion

The pixel decoder utilizes the off-the-shelf multi-scale deformable attention Transformer (MSDeformAttn) [46] with the default setting in [3, 46]. By using $\mathbf{X}^{'}$ (in Sec. 3.1) with a $1/32$ resolution as input, every 6 MSDeformAttn layers are taken to progressively up-sample the image features into $1/32$, $1/16$, $1/8$, and $1/4$, respectively. The decoded $1/32$, $1/16$, $1/8$ feature maps are denoted as $\mathbf{F}^{\times 32}$, $\mathbf{F}^{\times 16}$, $\mathbf{F}^{\times 8}$, respectively. The $1/4$ resolution feature map is directly utilized for per-pixel embedding.

Assume $\mathbf{F}^{\times 32}$, $\mathbf{F}^{\times 16}$ and $\mathbf{F}^{\times 8}$ correspond to the key and value of $\{\mathbf{V}^{\times 32}, \mathbf{K}^{\times 32}\}$, $\{\mathbf{V}^{\times 16}, \mathbf{K}^{\times 16}\}$ and $\{\mathbf{V}^{\times 8}, \mathbf{K}^{\times 8}\}$, respectively. Following Eqs. 4, 5, 6 7, the normalized key and value is generated by $\{\mathbf{V}^{'\times 32}, \mathbf{K}^{'\times 32}\}$, $\{\mathbf{V}^{'\times 16}, \mathbf{K}^{'\times 16}\}$ and $\{\mathbf{V}^{'\times 8}, \mathbf{K}^{'\times 8}\}$, respectively.

The Transformer decoder consists of 9 layers, where $L = 0, 1, \cdots, 8$. The feature propagation follows the procedure of Eq. 3, but each layer is fed into the normalized key and query. The leverage of the multi-scale image features is through a progressive manner. The image features from $\times 32$, $\times 16$ and $\times 8$ are subsequently embedded into the Transformer decoder in an end-to-end manner, given by

$$\mathbf{X}_i = \mathrm{softmax}(\mathcal{M}_{i-1} + \mathbf{Q}_{i-1}\mathbf{K}^{'\times 32\mathsf{T}})\mathbf{V}^{'\times 32} + \mathbf{X}_{i-1}, \quad (9)$$

$$\mathbf{X}_j = \mathrm{softmax}(\mathcal{M}_{j-1} + \mathbf{Q}_{j-1}\mathbf{K}^{'\times 16\mathsf{T}})\mathbf{V}^{'\times 16} + \mathbf{X}_{j-1}, \quad (10)$$

$$\mathbf{X}_k = \mathrm{softmax}(\mathcal{M}_{k-1} + \mathbf{Q}_{k-1}\mathbf{K}^{'\times 8\mathsf{T}})\mathbf{V}^{'\times 8} + \mathbf{X}_{k-1}, \quad (11)$$

where $i = 1, 4, 7$, and $j = 2, 5, 8$ and $k = 3, 6, 9$.

3.4. Network Architecture and Implementation

The overall framework is shown in Fig. 3. As our method intends to exploit the possibility of vision Transformer (ViT) for this task, we use the Mask2Former [3] as the feature extractor with a backbone of Swin-Transformer [17]. The pre-trained model from ImageNet is utilized as the initial weight parameters. The $1/4$ resolution feature map is fused with the features from the Transformer decoder for dense prediction.

All experiments are conducted on a work station with 64GB memory, an Intel® Core™ i7-10700K CPU and two GeForce RTX 2080 Ti GPUs. The batch size is set 2 per GPU. The Adam optimizer is used with an initial learning rate of 1×10^{-4}. The weight decay is set 0.05. The training terminates after 50 epochs.

Following the default setting of Mask2Former [3], the final loss function $\mathcal{L}$ is a linear combination of binary cross-entropy loss $\mathcal{L}_{ce}$, dice loss $\mathcal{L}_{dice}$, and the classification loss $\mathcal{L}_{cls}$, given by

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{cls}\mathcal{L}_{cls}, \quad (12)$$

where the hyper-parameters $\lambda_{ce} = \lambda dice = 5.0, \lambda_{cls} = 2.0$ keep the default as Mask2Former without any tuning.

4. Experiments

4.1. Dataset & Evaluation Protocols

Considering existing methods of domain generalization segmentation for driving-scenes, five semantic segmentation datasets are used in our experiments.

Specifically, CityScapes [5] provides 2,975 and 500 well-annotated samples for training and validation, respectively. These driving-scenes are captured in tens of Germany cities with a high resolution of 2048×1024.

BDD-100K [38] also provides diverse urban driving scenes with a resolution of 1280×720. 7,000 and 1,000 well-annotated samples are provided for training and validation of semantic segmentation, respectively.

Mapillary [21] is also a real-world large-scale semantic segmentation dataset with 25,000 samples from a variety of samples.

SYNTHIA [31] is large-scale synthetic dataset, and provides 9,400 images with a high resolution of 1280×760 for semantic segmentation.

GTA5 [30] a synthetic semantic segmentation dataset rendered by the GTAV game engine. It provides 24,966 simulated urban-street samples with a resolution of 1914×1052.

For clarity, we use C, B, M, S and G to denote these five datasets, respectively.

Following prior urban-scene domain generalized semantic segmentation works [4, 22–24], the segmentation model is trained on only one dataset as the source domain, and is validated on the rest of the four datasets as the target domain. Two settings include: 1) G to C, B, M, S; and 2) C to B, M, G, S. mIoU (in percentage %) is used as the validation metric.

For fair comparison between each CNN based domain generalized segmentation methods, all the reported performance is directly cited from prior works under the ResNet-50 backbone [4, 22–24].

4.2. Comparison with State-of-the-art

CityScapes Source Domain Table 1 reports the performance of the proposed INFormer on target domain of B,

Method	Proc. & year	Trained on Cityscapes (C)			
		→ B	→ M	→ G	→ S
IBN [22]	ECCV2018	48.56	57.04	45.06	26.14
IW [23]	CVPR2019	48.49	55.82	44.87	26.10
Iternorm [12]	CVPR2019	49.23	56.26	45.73	25.98
DRPC [40]	ICCV2019	49.86	56.34	45.62	26.58
ISW [4]	CVPR2021	50.73	58.64	45.00	26.20
GTR [25]	TIP2021	50.75	57.16	45.79	26.47
DIRL [37]	AAAI2022	51.80	-	46.52	26.50
SHADE [42]	ECCV2022	50.95	60.67	48.61	27.62
SAW [24]	CVPR2022	52.95	59.81	47.28	28.32
WildNet [15]	CVPR2022	50.94	58.79	47.01	27.95
AdvStyle [43]	NIPS2022	-	-	-	-
Ours	2023	**58.50**	**71.61**	**56.43**	**41.11**

Table 1. Performance comparison of the proposed INFormer and other CNN based domain generalization segmentation methods under the setting of: C→{B, M, G, S}. Evaluation metric mIoU is given in percentage (%).

M, G and S, respectively, after trained on the source domain C. The proposed INFormer shows a performance gain of 5.55%, 10.94%, 7.82% and 12.79% mIoU on the B, M, G and S dataset against the state-of-the-art CNN based method.

As the BDD100K dataset contains many nigh-time urban-street images, it is particularly challenging for existing urban-scene domain generalized segmentation methods. Still, a performance gain of 5.55% is obtained by the proposed INFormer.

GTA5 Source Domain Table 2 reports the performance of the proposed INFormer on target domain of C, B, M and S, respectively, after trained on the source domain G. The proposed INFormer shows a performance improvement of 9.87%, 10.70%, 13.58% and 12.08% against the existing state-of-the-art CNN based method on the C, B, M and S dataset, respectively.

These outcomes further demonstrate the strong feature generalization nature of the proposed INFormer. The training domain GTA5 is a synthetic segmentation dataset. Even if trained on the synthetic data, the proposed INFormer still shows the strongest performance on multiple real-world datasets, such as cityscapes (C) and BDD-100K (B).

Parameter Number & GFLOPs Under the C→S setting, the parameter number (denoted as Para. num.) and GFLOPs of existing CNN based domain generalized methods are further compared with the proposed INFormer.

It can be derived from Table 3 that, although the use of ViT as feature extractor doubles the parameter number and halves the GFLOPs, it leads to a mIoU performance gain of 14.61% and 14.91% against DIRL [37] and ISW [4], respectively.

Method	Proc.& year	Trained on GTA5 (G)			
		→ C	→ B	→ M	→ S
IBN [22]	ECCV2018	33.85	32.30	37.75	27.90
DRPC [40]	ICCV2019	37.42	32.14	34.12	28.06
IW [23]	CVPR2019	29.91	27.48	29.71	27.61
Iternorm [12]	CVPR2019	31.81	32.70	33.88	27.07
ISW [4]	CVPR2021	36.58	35.20	40.33	28.30
GTR [25]	TIP2021	37.53	33.75	34.52	28.17
DIRL [37]	AAAI2022	41.04	39.15	41.60	-
SHADE [42]	ECCV2022	44.65	39.28	43.34	-
SAW [24]	CVPR2022	39.75	37.34	41.86	30.79
WildNet [15]	CVPR2022	44.62	38.42	46.09	31.34
AdvStyle [43]	NIPS2022	39.62	35.54	37.00	-
Ours	2023	**54.52**	**49.98**	**59.67**	**43.42**

Table 2. Performance comparison of the proposed INFormer and other CNN based domain generalization segmentation under the setting of: G→{C, B, M, S}. Evaluation metric mIoU is presented in percentage (%).

Method	Backbone	GFLOPs	Para. num.	mIoU (%)
IBN [22]	ResNet-50	554.31	45.08	26.14
IW [23]		554.31	45.08	26.10
ISW [4]		554.31	45.08	26.20
DIRL [37]		554.98	45.41	26.50
INFormer	Mask2Former	223.37	107.21	**41.11**

Table 3. Comparison of parameter number, GFLOPs and mIoU of the proposed INFormer with some existing CNN based domain generalized methods. all the statistics are reported under the setting of: C→{B, M, G, S}.

4.3. Ablation Studies

Table 4 provides an ablation study on each component of the proposed INFormer. On top of the segmentation network Mask2Former [3], two components are considered, namely, instance normalization in the encoding stage (denoted as EN) and instance normalization in the decoding stage (denoted as DN), respectively.

The EN component leads to a performance gain of 1.35%, 1.85%, 1.02% and 1.97% on B, M, G and S target domain, respectively. The DN component leads to a performance gain of 2.58%, 4.07%, 1.37% and 2.89% on B, M, G and S target domain, respectively. The normalization in the decoding stage is more significant than the encoding stage.

4.4. Visualization

Some segmentation prediction results on the target domains are shown in Fig. 4. Compared to the CNN based domain generalized segmentation methods, the proposed INFormer shows a better segmentation prediction, especially in terms of the completeness of objects. Hence, the ViT based framework has promising application value in the domain generalized segmentation task.

These outcomes indicate that, for safety-crucial applications such as autonomous driving, when deploying domain

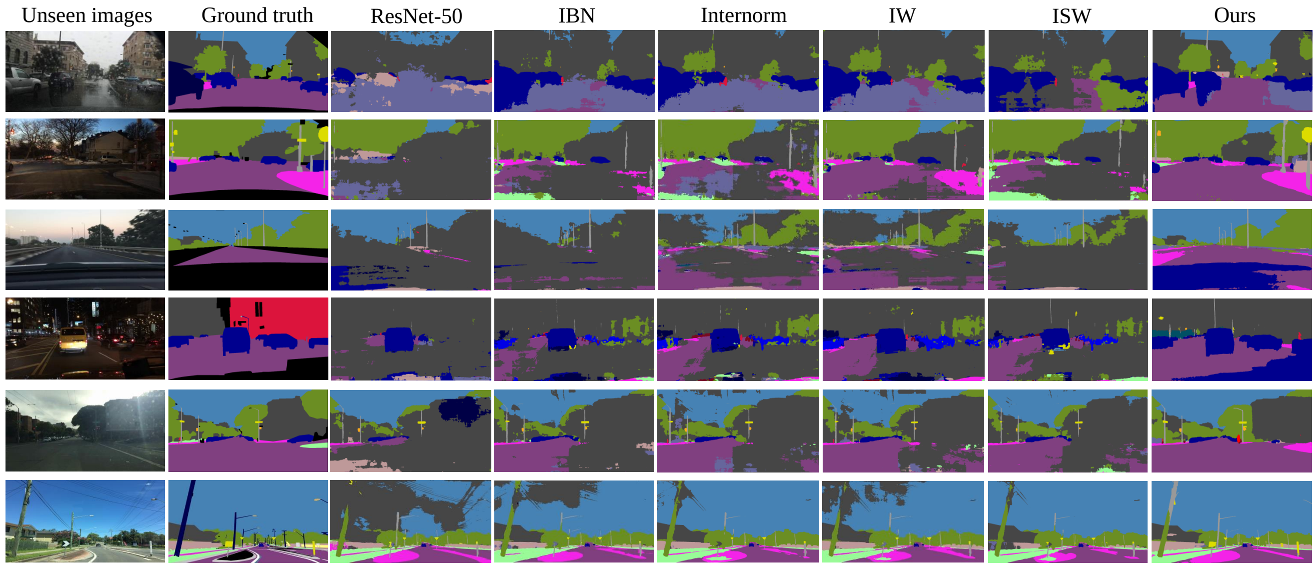

Figure 4. Segmentation prediction of existing CNN based domain generalized semantic segmentation methods (IBN [22], IW [23], Iternorm [12], ISW [4]) and the proposed INFormer on the images from unseen target domains.

Component			Trained on CityScapes (C): mIoU (%)			
Mask2Former	EN	DN	→ B	→ M	→ G	→ S
✓			55.43	66.12	55.05	38.19
✓	✓		56.78	67.97	56.07	40.16
✓		✓	58.01	70.19	56.42	41.08
✓	✓	✓	**58.50**	**71.61**	**56.43**	**41.11**

Table 4. Ablation studies on each component of the proposed INFormer under the setting of: C→{B, M, G, S}. EN / DN: instance normalization in the Transformer encoding / decoding stage. Evaluation metric mIoU is presented in percentage (%).

generalized segmentation algorithms, ViT based methods are preferred.

5. Conclusion & Limitation Discussion

In this paper, we investigate the possibility to adapt the vision Transformer for the task of domain generalized semantic segmentation. An instance normalization Transformer (INFormer) is proposed for this task. The key idea is that the feature embeddings in ViT are normalized during both the encoding and decoding stage in a progressive manner. Extensive experiments on multiple domain generalized segmentation settings show the superior performance of the proposed INFormer against existing CNN based domain generalized segmentation methods. Moreover, the visualization also shows the superior qualitative inference of the proposed INFormer than existing methods.

Limitation discussion. As the feature extraction pipeline of ViT and CNN is quite different, the proposed progressive normalization strategy needs to calculate the key and value for the self-attention mechanism. Thus, it is not directly applicable to existing CNN based domain generalized segmentation pipelines. Nevertheless, its effectiveness against the baseline is demonstrated by the ablation study. Performance superiority against existing CNN based domain generalized segmentation methods is shown.

References

[1] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020. 2

[2] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17653–17662, 2022. 1

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1, 3, 4, 5

[4] S. Choi, S. Jung, H. Yun, J. Kim, S. Kim, and J. Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11580–11590, 2021. 1, 2, 4, 5, 6

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

ings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016. 2, 4

[6] Carlos A Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21383–21392, 2022. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1

[8] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[9] Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, et al. Unsupervised domain generalization by learning a bridge across domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5280–5290, 2022. 2

[10] Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 2

[11] L. Huang, Y. Zhou, F. Zhu, L. Liu, and L. Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4874–4883, 2019. 1, 2

[12] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 4874–4883, 2019. 5, 6

[13] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4350–4360, 2022. 2

[14] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888, 2020. 2

[15] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 2, 5

[16] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. 1

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10012–10022, 2021. 1, 4

[18] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 2

[19] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11749–11756, 2020. 2

[20] M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3011, 2022. 1

[21] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 2, 4

[22] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 1, 2, 4, 5, 6

[23] X. Pan, X. Zhan, J. Shi, X. Tang, and P. Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1863–1871, 2019. 1, 2, 4, 5, 6

[24] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2594–2605, 2022. 1, 2, 4, 5

[25] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30:6594–6608, 2021. 2, 5

[26] Xi Peng, Fengchun Qiao, and Long Zhao. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[27] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, pages 7728–7738. PMLR, 2020. 2

[28] Fabrizio J Piva, Daan de Geus, and Gijs Dubbelman. Empirical generalization study: Unsupervised domain adaptation vs. domain generalization methods for semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 499–508, 2023. 2

[29] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 2

[30] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 2, 4

[31] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2, 4

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10765–10775, 2021. 1

[33] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115, 2023. 2

[34] Gabriel Tjio, Ping Liu, Joey Tianyi Zhou, and Rick Siow Mong Goh. Adversarial semantic hallucination for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 318–327, 2022. 2

[35] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020. 2

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 3

[37] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2884–2892, 2022. 1, 2, 5

[38] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 2, 4

[39] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019. 2

[40] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2100–2110, 2019. 2, 5

[41] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33:16096–16107, 2020. 2

[42] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 535–552. Springer, 2022. 2, 5

[43] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In *Advances in Neural Information Processing Systems*, 2022. 2, 5

[44] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[45] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020. 2

[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 4