Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Learning generalized visual relations for domain generalization semantic segmentation

Zijun Li [b], Muxin Liao [a],*

[a] School of Computer Science and Engineering, Jiangxi Agricultural University, Lushan South Avenue, Nanchang, 330000, Jiangxi Province, China
[b] School of Public Policy and Administration, Chongqing University, 400044, Chongqing, China

ARTICLE INFO

Keywords:
Domain generalization
Domain-invariant inter-class relation
Intra-class relation
Semantic segmentation

ABSTRACT

Domain generalization semantic segmentation (DGSS) methods aim to generalize well on out-of-distribution scenes, which is crucial for real-world applications. The key to DGSS methods is to improve the generalization ability and class discriminability in unseen domains. For one thing, visual relations among classes provide prior knowledge about the real world, which contains domain-invariant information. For another, visual relations within classes are exploited to learn their respective internal structures, which provide class-discriminability information. Based on this, a generalized visual relations learning (GVRL) approach is proposed, which contains Domain-invariant IntEr-class Relation Learning (DIERL) and InTrA-class Relation Learning (ITARL). For the DIERL, first, a prototype of each class, which is consistent across domains, is constructed to learn domain-invariant class-specific features for mining inter-class relations. Second, the consistency constraint of the inter-class relations between the source and augmented domains is proposed to learn domain-invariant inter-class relations. For the ITARL, since there is no obvious intra-class relation in stuff classes (e.g., road, sky, and building), only the intra-class relation of instance classes (e.g., person, car, and rider) is learned to enhance the diversity of features for improving the class discriminability. Extensive experiments demonstrate that our approach achieves superior average performance over current approaches on three DGSS benchmarks, e.g., the proposed approach based on the ResNet-50 backbone achieves average performances of 44.8%, 38.7%, and 45.6% in terms of mIoU, improving by 0.9%, 1.1%, and 1.6% over current approaches.

1. Introduction

Supervised-based semantic segmentation has achieved remarkable progress in multiple real-world applications, such as autonomous driving (González-Collazo, Balado, González, & Nurunnabi, 2023; Zeng et al., 2024), remote sensing (Barbato, Piccoli, & Napoletano, 2024; Fan, Zhou, Qian and Yan, 2024), and medical diagnosing (Paithane & Kakarwal, 2023; Suo, Zhou, Hu, Zhang, & Gao, 2024). Existing supervised-based semantic segmentation methods (Elhassan, Huang, Yang, & Munea, 2021; Xian et al., 2022) are trained and tested on the independent and identically distributed data. However, the performance of these methods drops significantly on the out-of-distribution testing data when the training data (source domain) and testing data (target domain) are from different domains. To solve this problem, unsupervised domain adaptation semantic segmentation (UDASS) methods (Dong, Niu, Gao, Li, & Shao, 2024; Hong, Kumar, & Lee, 2024; Liao et al., 2022, 2024b, 2024c) are proposed to reduce the distribution shift. The goal of UDASS methods is to learn domain-invariant features from the labeled source domain and the unlabeled target domain. Since UDASS methods require to access the unlabeled target domain while it is quite difficult to create a target domain that covers all real unseen scenes, the performance of UDASS methods still degrades when testing out-of-distribution scenes.

To tackle this challenge, domain generalization semantic segmentation (DGSS) methods (Liao et al., 2023a, 2023b, 2024a) are proposed. Since the training of DGSS cannot access target domain data, the goal of DGSS is to learn domain-invariant features from the labeled source domain and generalize well in unseen domains. Thus, DGSS can be seen as a special issue of UDASS. Existing DGSS methods (Choi et al., 2021; Pan, Luo, Shi, & Tang, 2018; Pan, Zhan, Shi, Tang, & Luo, 2019; Xu et al., 2022) aim to learn domain-invariant features by removing domain-specific information for improving the generalization ability. Although these methods achieve better performance on unseen domains to some extent, they ignore the learning of visual relations. For one thing, visual relations among classes provide prior knowledge about the real world, which contains domain-invariant information. For another, visual relations within classes are exploited to learn their respective

* Corresponding author.
E-mail addresses: 20215858@stu.cqu.edu.cn (Z. Li), liaomx@jxau.edu.cn (M. Liao).

https://doi.org/10.1016/j.eswa.2024.126217
Received 9 March 2024; Received in revised form 11 October 2024; Accepted 17 December 2024
Available online 24 December 2024

internal structures, which provide class-discriminability information. Thus, we intuitively argue that learning generalized visual relations (including inter-class and intra-class relations) can help improve the generalization ability and class discriminability. In particular, for the semantic segmentation task, the inter-class relations mean the relations between different classes within an image. For example, in urban scenes, vehicles are located on roads, pedestrians are typically near sidewalks or crosswalks, and trees are often near the edges of roads. Learning inter-class relations is beneficial for mining the contextual information of objects which better differentiates between classes for making more accurate segmentation results. The intra-class relations mean the relations within a class, such as the relations between the tires and bodywork of the car class, the relations between the person and bicycle of the rider class, etc. Learning intra-class relations is beneficial for capturing intra-class variations that allow the model to segment diverse objects that differ in style but belong to the same class in unseen domains.

Based on this, a novel approach, named generalized visual relation learning (GVRL), is proposed to learn the domain-invariant inter-class visual relations and intra-class visual relations for DGSS. Specifically, the GVRL contains domain-invariant inter-class relation learning (DIERL) and intra-class relation learning (ITARL). For the DIERL, the domain-invariant inter-class relations of each class are learned from two aspects. For one thing, since the class-specific features are sensitive to the distribution shift, the inter-class relations learned by these class-specific features may be unstable. Thus, a prototype of each class, which is consistent across domains, is constructed to learn domain-invariant class-specific features for mining stable inter-class relations. For another, to further ensure the learned inter-class relations are domain-invariant when the distribution shift problem happens, the consistency constraint of the inter-class relations between the source and augmented domains is proposed. For the ITARL, since there is no obvious intra-class relation in stuff classes (e.g., road, sky, and building), only the intra-class relation of instance classes (e.g., person, car, and rider) is learned to enhance the diversity of features for improving the class discriminability. Extensive experiments demonstrate that the GVRL achieves superior performance over current approaches on domain generalization segmentation tasks. The contributions are summarized as follows:

1. This paper proposes a generalized visual relation learning (GVRL) approach for domain generalization semantic segmentation. The GVRL contains domain-invariant inter-class relation learning (DIERL) and intra-class relation learning (ITARL) to simultaneously improve the generalization ability and the class discriminability.
2. The DIERL is proposed to learn domain-invariant inter-class relations by constructing a prototype of each class and using the consistency constraint of the inter-class relations between the source and augmented domains for improving the generalization ability.
3. The ITARL aims to learn the richer representations of instance classes for improving class discriminability since instance classes contain more obvious intra-class structures than stuff classes.
4. The GVRL achieves superior performance against the state-of-the-art methods on multiple challenging tasks for domain generalization semantic segmentation.

The rest of the paper is organized as follows. In Section 2, related work is reviewed. In Section 3, the unified framework, domain-invariant inter-class relation learning, and intra-class relation learning are introduced. In Section 4, extensive experiments are conducted, and some visualizations are analyzed accordingly. Finally, some concluding remarks are offered in Section 5.

2. Related work

In this section, we summarize some recent related methods, including domain generalization and visual relation learning for semantic segmentation.

2.1. Domain generalization methods for semantic segmentation

Domain generalization methods (Asutkar, Chalke, Shivgan, & Tallur, 2023; Fan, Chen et al., 2024; Guo et al., 2024) aim to learn a model from single or multiple source domains that can perform well on out-of-distribution scenes. Domain generalization can be seen as a special issue of domain adaptation. On one hand, similar to domain adaptation methods, the goal of domain generalization methods is also the learning of domain-invariant features for improving the generalization ability. On the other hand, different from domain adaptation methods, the training of domain generalization methods cannot access target domain data, which means domain generalization is more challenging than domain adaptation. Domain generalization methods are divided into three categories, including data manipulation, representation learning, and learning strategy. In this subsection, we mainly review and discuss domain generalization methods on semantic segmentation.

Domain generalization methods based on data manipulation are mainly divided into two categories: domain randomization and feature mixup. For example, Huang, Guan, Xiao, and Lu (2021), Peng, Lei, Liu, Zhang, and Liu (2021), Tjio, Liu, Zhou, and Goh (2022) and Yue et al. (2019) are proposed to randomize the styles of source domain images by using domain randomization for improving the generalization ability on different styles. Lee, Seong, Lee, and Kim (2022) propose content extension learning by using feature mixup to mix the source domain content and wild content for alleviating the source-content overfitting problems. Liao et al. (2024a) propose class-wise and element-wise prototype augmentation strategies to learn domain-invariant representations from augmented prototypes.

Domain generalization methods based on representation learning often use feature normalization (Pan et al., 2018; Ulyanov, Vedaldi, & Lempitsky, 2017) or whitening (Pan et al., 2019) methods to learn domain-invariant features. Choi et al. (2021) proposes an instance selective whitening strategy to normalize domain-specific features for learning domain-invariant features. Xu et al. (2022) propose a feature whitening strategy to remove the feature correlations which are sensitive to the domain-specific style. Peng, Lei, Hayat, Guo, and Li (2022) propose semantic-aware normalization and semantic-aware whitening to encourage both intra-class compactness and inter-class separability. Liao et al. (2023b) proposes to aggregate the edge and semantic layout information to guide learning domain-invariant content information. Huang et al. (2023) proposes to build a better feature space via style projection and semantic clustering. Zhang, Tian, Liao, Hua et al. (2023) proposes to learn shape-invariant representation for improving the generalization ability. Liao et al. (2023a) proposes to calibrate the uncertainty of prototypes during the prototypical contrastive learning for learning domain-invariant features. Zhang, Tian, Liao, Zhang et al. (2023) proposes to learn fine-grained domain-invariant features via self-supervision for improving the generalization ability. Yang, Gu, and Sun (2023) propose to use a self-supervised source domain projection strategy to transform the style of unseen domains into the style of the source domain and then learn domain-invariant features by using contrastive learning.

Moreover, except for data manipulation and representation learning, other domain generalization methods use different learning strategies to improve the generalization ability, such as meta-learning strategy (Kim, Lee, Park, Min, & Sohn, 2022) and adversarial learning strategy (Zhao, Zhong, Zhao, Sebe, & Lee, 2022).

The proposed approach falls into the second category, which is based on representation learning. Unlike existing methods based on representation learning, which learns domain-invariant representation by eliminating domain-specific information or keeping representation consistency with the same contents in different styles, these methods ignore the learning of domain-invariant visual relations to enhancing domain-invariant information and preserving class-discriminative information. In this paper, the proposed approach aims to improve generalization ability and class discriminability by learning domain-invariant inter-class and intra-class relations.

2.2. Visual relation learning for semantic segmentation

Visual relation learning is widely used in many tasks, such as semantic segmentation (Ma, Guo, Liu, Lei, & Wen, 2020; Niu et al., 2021; Qi, Liao, Jia, Fidler, & Urtasun, 2017), panoptic segmentation (Borse et al., 2022; Wu et al., 2020), videos processing (Baradel, Neverova, Wolf, Mille, & Mori, 2018), visual question answering (Nguyen et al., 2022; Wang, Bao, & Xu, 2021; Zhang, Jiang, & Zhao, 2021), and pedestrian intent prediction (Liu et al., 2020). We mainly review and discuss the progress of visual relation learning in the semantic segmentation task. Visual relation learning methods are basically divided into two categories: self-attention mechanism based method (Vaswani et al., 2017) and graph convolution network based ones (Chen et al., 2019).

The self-attention mechanism (Vaswani et al., 2017) is mostly utilized to capture the visual dependency relations of each pixel by using context aggregation. Wang, Girshick, Gupta, and He (2018) proposes a non-local network to use the self-attention mechanism for modeling the long-range context in an image. Since Wang et al. (2018) only focuses on context aggregation in the spatial dimensions, Fu et al. (2019) proposes a position attention module and a channel attention module to respectively capture the position and channel dependencies relations of each pixel from spatial and channel dimensions. Moreover, many works propose the different variants of the non-local network (Wang et al., 2018) to capture richer contextual information. Tang et al. (2022) proposes a multi-scale context aggregation module to avoid the drawback of losing local contextual continuity. Ding, Shen, Zeng, and Peng (2022) proposes a class-wise semantic attention module to balance the non-local contextual dependencies and the local consistency with a class-wise weight. However, when these methods compute the attention map with the high resolution of the input feature maps, the computational cost of these methods became higher. To alleviate this problem, an efficient non-local attention module has been proposed. Huang et al. (2019) proposes a criss-cross network by using a recurrent criss-cross context aggregation operation to capture the full-image dependencies from all pixels and reduce the computational cost. He, Liu, Wang, and Lu (2022) proposes an efficient sampling-based attention network that contains a deterministic sampling-based attention module to sample features from a local region for selecting interesting information.

Different from the self-attention mechanism which captures the visual relations of each pixel, the graph convolution network (Chen et al., 2019) is always leveraged to capture the visual relations between objects or regions of the image. Chen et al. (2019) proposes a novel relations reasoning approach by using graph convolution to capture relations between arbitrary regions over the full input space. Xie, Liu, Xiong, and Shao (2021) proposes an end-to-end scale-aware graph neural network to reason the cross-scale relations among the support-query images for few-shot semantic segmentation. Zhou et al. (2021) proposes a novel group-wise learning framework to explicitly encode semantic dependencies in a group of images by using a graph neural network. Moreover, to reduce the computational cost of graph reasoning, Li et al. (2021) proposes an efficient framework called squeeze reasoning to capture contextual relations by using squeezed features to build the node graph. Different from these methods which capture the visual relations between regions of the image, Hu et al. (2020) proposes a class-wise dynamic graph convolution module to reason visual relations among pixels in the same class. Liu, Schonfeld, and Tang (2021) proposes a novel dependency network by using a graph to reason inter-class dependency relations and using two convolution layers to reason intra-class dependency relations.

In particular, the idea of Liu et al. (2021) which proposes to learn intra-class and inter-class relations is similar to our approach. In Liu et al. (2021), for intra-class relations learning, the features extracted from the backbone is decoupled into class-specific features by using group convolutions. Then, the class-specific features are used for inter-class relations learning. For inter-class relations learning, first, the dependency relations among different classes are captured by using a dependence graph and then the spatial and semantic learning of inter-class dependence relations are implemented by using group-weighted convolution. Thus, the whole process is step-by-step.

Compared with (Liu et al., 2021), the GVRL has two differences, including the motivation for visual relations learning and the fashion of visual relations learning. For the motivation of two approaches, the GVRL aims to learn domain-invariant inter-class relations for DGSS while (Liu et al., 2021) ignores the domain-specific information causing distribution shift of the inter-class relations since (Liu et al., 2021) is utilized for supervised semantic segmentation. For the fashion of visual relations learning, the process of inter-class and intra-class relations learning is parallel in the GVRL. Specifically, the inter-class relations and intra-class relations are learned by domain-invariant class-specific features which are obtained by constructing a prototype of each class. Moreover, since there is no obvious intra-class relation in stuff classes (e.g., road, sky, and building), only the intra-class relation of instance classes (e.g., person, car, and rider) is learned.

3. Approach

In this section, we first give an overview of the GVRL and sequentially introduce the DIERL, the ITARL, and our total training loss.

3.1. Overview

The generalized visual relation learning (GVRL) contains Domain-Invariant intEr-class Relations Learning (DIERL) and InTrA-class Relation Learning (ITARL) which aim to respectively learn domain-invariant inter-class relations and intra-class relations for domain generalization semantic segmentation. The flowchart of the GVRL is illustrated in Fig. 1. Specifically, to eliminate distribution-sensitive information during the learning of domain-invariant inter-class relations, first, data augmentation strategies, such as color jitter, Gaussian blur, flip horizontal, and flip vertical, are applied to the source domain images X^s for generating the augmented domain images X^{aug}, which are viewed as unseen domain images. Second, the features Z_x^s and $Z_x^{aug} \in \mathbb{R}^{N\times H\times W}$ extracted from the source and augmented domains by using the backbone are used to obtain the class-specific features Z_c^s and $Z_c^{aug} \in \mathbb{R}^{C\times H\times W}$, which are denoted as follows:

$$Z_c^s = \frac{\sum_{n=1}^{N}\sum_{i=1}^{H}\sum_{j=1}^{W} Z^s_{x,n,i,j}\,\mathbb{1}[Y_{n,i,j}=c]}{\sum_{n=1}^{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbb{1}[Y_{n,i,j}=c]} \tag{1}$$

$$Z_c^{aug} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{H}\sum_{j=1}^{W} Z^{aug}_{x,n,i,j}\,\mathbb{1}[Y_{n,i,j}=c]}{\sum_{n=1}^{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbb{1}[Y_{n,i,j}=c]} \tag{2}$$

where Y is the ground truth, c denotes the classes, N is the number of the Z_x^s and Z_x^{aug}, H and W respectively denote the height and the width of the Z_x^s and Z_x^{aug}, $\mathbb{1}[Y_{n,i,j}=c]$ is an indicator function, which equals to 1 if $Y_{n,i,j} = c$ and 0 otherwise. Then, the Z_c^s and Z_c^{aug} are utilized to initial the prototype Z_p, which is denoted as follows:

$$Z_p = \frac{1}{2}(Z_c^s + Z_c^{aug}) \tag{3}$$

where the prototype Z_p is updated by the Z_c^s and Z_c^{aug} in every iteration, which is formulated as follows:

$$Z_p = mZ_p + (1-m)\frac{1}{2}(Z_c^s + Z_c^{aug}) \tag{4}$$

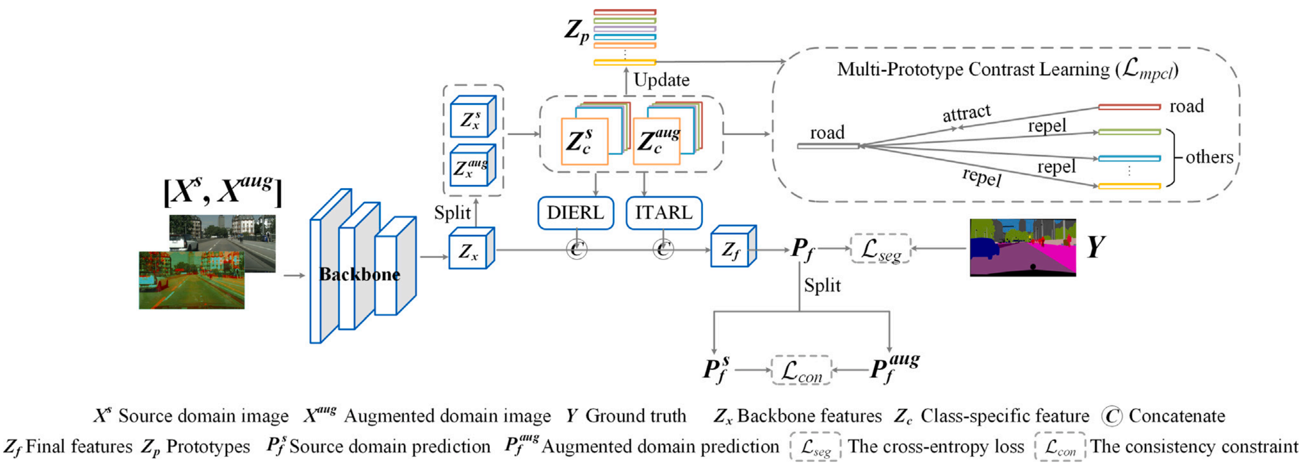

Z_p
Update
Multi-Prototype Contrast Learning ($\mathcal{L}_{mpcl}$)
road
attract
repel
road
repel
others
repel
$[X^s, X^{aug}]$
Backbone
Split
Z_x^s
Z_x^{aug}
Z_c^s
Z_c^{aug}
DIERL
ITARL
Z_x
C
C
Z_f
P_f
$\mathcal{L}_{seg}$
Y
Split
P_f^s
$\mathcal{L}_{con}$
P_f^{aug}
X^s Source domain image
X^{aug} Augmented domain image
Y Ground truth
Z_x Backbone features
Z_c Class-specific feature
C Concatenate
Z_f Final features
Z_p Prototypes
P_f^s Source domain prediction
P_f^{aug} Augmented domain prediction
$\mathcal{L}_{seg}$ The cross-entropy loss
$\mathcal{L}_{con}$ The consistency constraint

Fig. 1. The proposed approach for domain generalization semantic segmentation.

where m is a trade-off parameter. To learn domain-invariant class-specific features Z_c^s and Z_c^{aug}, the multi-prototype contrast learning loss $\mathcal{L}_{mpcl}$ between the prototype Z_p and the class-specific features Z_c^s and Z_c^{aug} is computed, which encourages the class-specific features Z_c^s and Z_c^{aug} to be closer to their assigned prototypes. The $\mathcal{L}_{mpcl}$ is denoted as follows:

$$\mathcal{L}_{mpcl} = -\sum_{i=1}^{C} Y_i \log \frac{\frac{1}{2}(exp(Z_{p^i} \cdot Z^s_{c^i}/\tau) + exp(Z_{p^i} \cdot Z^{aug}_{c^i}/\tau))}{\frac{1}{2}(\sum_{i \neq k}^{C} exp(Z_{p^k} \cdot Z^s_{c^i}/\tau) + \sum_{i \neq k}^{C} exp(Z_{p^k} \cdot Z^{aug}_{c^i}/\tau))} \tag{5}$$

where i and k are the index of the class number. C is the number of classes. Z_{p^i} and Z_{p^k} denote the prototypes of ith and kth classes, respectively. $Z^s_{c^i}$ and $Z^{aug}_{c^i}$ denotes the class-specific features of ith class from the source and augmented domains, respectively. τ is the temperature parameter. Y is the ground truth. $\cdot$ is the dot product. $exp(\cdot)$ is used to compute the cosine similarity between two features. The value of the $\mathcal{L}_{mpcl}$ is low when the Z_c^s and Z_c^{aug} are similar to its prototype Z_p and dissimilar to all other prototypes. Thus, the $\mathcal{L}_{mpcl}$ is minimized to learn domain-invariant class-specific features which are used to learn domain-invariant inter-class relations Z_{inter} by the DIERL and intra-class relations Z_{intra} by the ITARL. Finally, the cross-entropy loss $\mathcal{L}_{seg}$ is optimized the Z_f for predicting more accurate segmentation maps P_f. The $\mathcal{L}_{seg}$ is denoted as follows:

$$\mathcal{L}_{seg}(P_f, Y) = -Y \cdot \log(P_f) \tag{6}$$

where Y is the ground truth, P_f is the predicted segmentation maps obtained by upsampling the Z_f. In particular, Z_f is obtained by concatenating the Z_x, Z_{inter}, and Z_{intra}. By minimizing the cross-entropy loss $\mathcal{L}_{seg}$, the inter-class and intra-class relations are prompted to work with the semantic segmentation.

Moreover, to keep the consistent prediction between the source and augmented domains, the consistency constraint between the P_f^s and P_f^{aug}, which are split by the P_f in the dimensionality of batch size, is computed, which is denoted as follows:

$$\mathcal{L}^p_{con}(P_f^s, P_f^{aug}) = 1 - \frac{P_f^s \cdot P_f^{aug}}{\|P_f^s\| \cdot \|P_f^{aug}\|} \tag{7}$$

where the $\mathcal{L}^p_{con}$ is computed by the cosine similarity. The consistent prediction is learned by minimizing the $\mathcal{L}^p_{con}$. In particular, since the X^s and X^{aug} are concatenated by the dimensionality of batch size to feed into the model for generating the predictions P_f, the P_f^s and P_f^{aug} can be generated by splitting the P_f by the dimensionality of batch size. In the following subsections, we sequentially introduce the DIERL, the ITARL, and our total training loss.

3.2. The domain-invariant inter-class relation learning

The goal of the domain-invariant inter-class relation learning (DIERL) is to learn domain-invariant inter-class relations. To achieve this goal, the prototype of each class and the consistency constraint of inter-class relations between the source and augmented domains are considered to be used. For one thing, since the prototype of the same class on different domains is supposed to share a high representational similarity, the prototypes are utilized to depict the same inter-class relations from different domains. For another, the consistency constraint, which is used to measure the stability of inter-class relations across domains, aims to learn the consistent inter-class relations between the source and augmented domains. Based on these motivations, the DIERL is proposed to learn domain-invariant inter-class relations. The process is divided into three stages, which is shown in Fig. 2.

After obtaining the class-specific features Z_c^s and Z_c^{aug}, the DIERL performs semantic reasoning to get the inter-class relations Z^s_{inter} and Z^{aug}_{inter} by putting Z_c^s and Z_c^{aug} into the graph neural network (GNN) (Han, Wang, Guo, Tang, & Wu, 2022), respectively. Specifically, to build the GNN, first, the class-specific features Z_c^s and Z_c^{aug} with $H \times W \times C$ are respectively divided into C vectors z_i^s and $z_i^{aug} \in \mathbb{R}^D$, where D equals $H \times W$, $i = 1, 2, \dots, C$, and C is the number of classes. The vector of each class (z_i^s or z_i^{aug}) is viewed as a set of unordered nodes of the graph which are denoted as $\mathcal{V} = \{v_1, v_2, \dots, v_C\}$, its K nearest neighbors $\mathcal{N}(v_i)$ are found and an edge e_{ji} is added for all nodes $v_j \in \mathcal{N}(v_i)$, where e_{ji} denotes the edge directed from v_j to v_i. Up to this point, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is obtained, where $\mathcal{E}$ denotes all edges. Relying on the edges $\mathcal{E}$, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$ is constructed, which is used to mask the insensitive inter-class weights matrix to set the zero value if two nodes are unconnected, which are denoted as follows:

$$\tilde{A}^s = \tilde{M}^s \cdot A \tag{8}$$

$$\tilde{A}^{aug} = \tilde{M}^{aug} \cdot A \tag{9}$$

where $\tilde{A}^s$ and $\tilde{A}^{aug}$ denote the masked insensitive inter-class weights matrices, which are utilized to weigh the vertex features during the graph convolution process. This is beneficial for learning domain-invariant inter-class relations. $\tilde{M}^s$ and $\tilde{M}^{aug}$ denote the insensitive inter-class weights matrices of the Z_c^s and Z_c^{aug}. To generate the insensitive inter-class weights matrices $\tilde{M}^s$ and $\tilde{M}^{aug}$, the inter-class weights of Z_c^s and Z_c^{aug} are computed, which are denoted as follows:

$$M^s = Z_c^s \times (Z_c^s)^T \tag{10}$$

$$M^{aug} = Z_c^{aug} \times (Z_c^{aug})^T \tag{11}$$

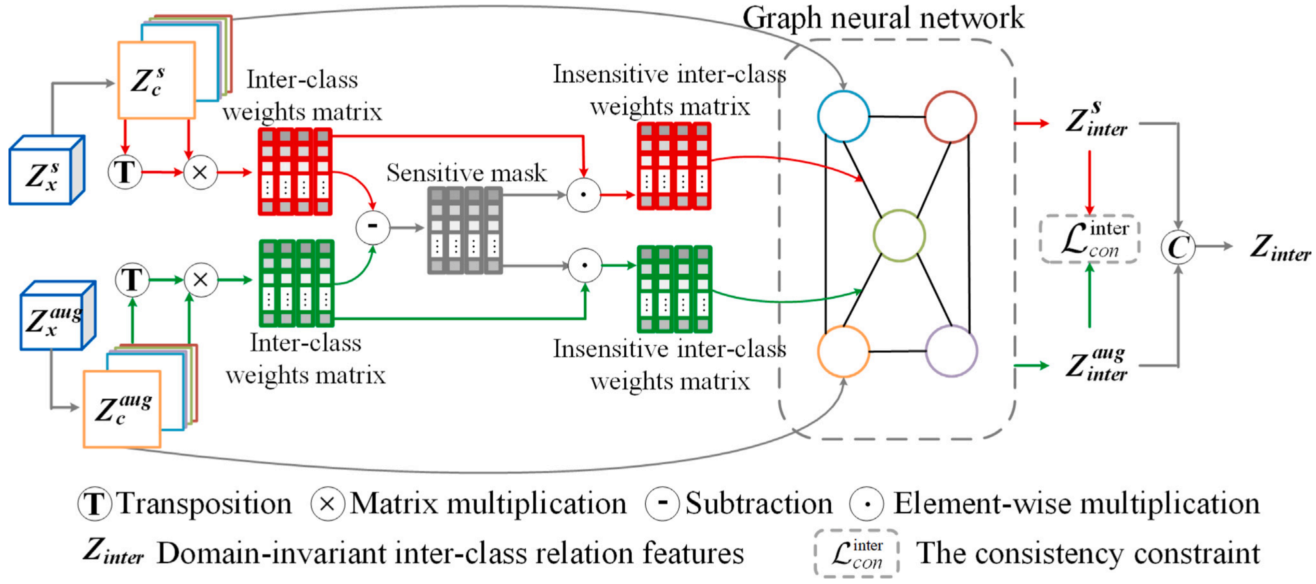

Graph neural network
Z_c^s
Z_x^s
Inter-class weights matrix
Insensitive inter-class weights matrix
Sensitive mask
Z^s_{inter}
$\mathcal{L}^{inter}_{con}$
Z_{inter}
Z^{aug}_{inter}
Z_x^{aug}
Z_c^{aug}
Inter-class weights matrix
Insensitive inter-class weights matrix
T Transposition
× Matrix multiplication
- Subtraction
· Element-wise multiplication
Z_{inter} Domain-invariant inter-class relation features
$\mathcal{L}^{inter}_{con}$ The consistency constraint

Fig. 2. The architecture of the DIERL.

where M^s and $M^{aug} \in \mathbb{R}^{C\times C}$. Since the inter-class weights may be sensitive to the distribution shift, to generate insensitive inter-class weights, first, the sensitive mask is obtained by computing the difference between the M^s and M^{aug}, which is denoted as follows:

$$U = \mathbf{1} - Softmax(abs(M^s - M^{aug})) \tag{12}$$

where $\mathbf{1}$ is an all-ones matrix. $Softmax(\cdot)$ is used to compute the probability matrix of the difference between the M^s and M^{aug}, where more smaller probability means more insensitive. To make the more insensitive value with a larger probability, subtract the probability matrix from the all-one matrix $\mathbf{1}$. Based on this, the uncertainty probability matrix U is obtained. In the uncertainty probability matrix U, a small probability represents that elements between the M^s and M^{aug} have big differences. In other words, a small probability means the corresponding element is more sensitive. Thus, the uncertainty probability matrix U is used to reweight the M^s and M^{aug} to reduce the sensitive weight for obtaining the insensitive inter-class weights $\tilde{M}^s$ and $\tilde{M}^{aug}$, which are denoted as follows:

$$\tilde{M}^s = M^s \cdot U \tag{13}$$

$$\tilde{M}^{aug} = M^{aug} \cdot U \tag{14}$$

Second, the graph $\mathcal{G}$ is put into a tiny pyramid graph neural network (Han et al., 2022) to learn inter-class relations Z^s_{inter} and Z^{aug}_{inter} by using a graph convolutional layer which can exchange information between nodes by aggregating features from its neighbor nodes. In particular, the tiny pyramid graph neural network (Han et al., 2022) is a vision graph neural network architecture,[1] which is used to extract graph-level features for visual tasks. The graph convolution operates $\mathcal{F}$ at lth layer can be denoted as follows:

$$\begin{aligned}\mathcal{G}_{l+1} &= \mathcal{F}(\mathcal{G}_l, \mathcal{W}_l)\\ &= Update(Aggregate(\mathcal{G}_l, \mathcal{W}_l^{agg}), \mathcal{W}_l^{update})\end{aligned} \tag{15}$$

where $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$ and $\mathcal{G}_{l+1} = (\mathcal{V}_{l+1}, \mathcal{E}_{l+1})$ are input and output graphs of lth graph convolutional layer, respectively. $\mathcal{W}_l^{agg}$ and $\mathcal{W}_l^{update}$ are the learnable weights of the aggregation and update functions in the graph convolutional layer, respectively, which can be denoted as follows:

$$z_{v_{l+1}} = h(z_{v_l}, g(z_{v_l}, \mathcal{N}(z_{v_l}), W_l^{agg}), W_l^{update}) \tag{16}$$

where $g(\cdot,\cdot,\cdot)$ is a vertex feature aggregation function and $h(\cdot,\cdot,\cdot)$ is a vertex feature update function. $\mathbf{Z}_{v_{l+1}}$ and $\mathbf{Z}_{v_l}$ are the vertex features at the lth and $l+1$th layers, respectively. $\mathcal{N}(\mathbf{Z}_{v_l})$ is the set of neighbor vertices of v at the lth layer. During these processes, to avoid the vertex features affecting by distribution shifts, the masked insensitive inter-class weights matrices are utilized to weigh the vertex features. Thus, the Eq. (16) can be re-formulated as follows:

$$z_{v_{l+1}} = h(\tilde{A}\cdot z_{v_l}, g(\tilde{A}\cdot z_{v_l}, \mathcal{N}(z_{v_l}), W_l^{agg}), W_l^{update}) \tag{17}$$

where $\tilde{A}$ denotes the masked insensitive inter-class weights matrices. For Z_c^s as the input of $\mathcal{G}$, the $\tilde{A}^s$ is used. For Z_c^{aug} as the input of $\mathcal{G}$, the $\tilde{A}^{aug}$ is used. Then, a feed-forward network (FFN) module is utilized on each vertex features. The FFN module consists of a simple multi-layer perceptron with two fully-connected layers, which can be denoted as follows:

$$z^{ffn}_{v_{l+1}} = \sigma(z_{v_{l+1}} W_{f1})W_{f2} + z_{v_{l+1}} \tag{18}$$

where W_{f1} and W_{f2} are the weights of fully-connected layers, σ is the activation function, e.g., ReLU and GeLU (Hendrycks & Gimpel, 2016). By stacking the graph convolutional layers and FFN module to a block, the GNN is built.

Finally, the inter-class relations Z^s_{inter} and Z^{aug}_{inter} can be generated by putting the class-specific features Z^s_c and Z^{aug}_c, which are respectively from the source and augmented domains, into the weight-shared GNN. To learn domain-invariant inter-class relations, the consistency constraint between the Z^s_{inter} and Z^{aug}_{inter} is leveraged, which is denoted as follows:

$$\mathcal{L}^{inter}_{con}(Z^s_{inter}, Z^{aug}_{inter}) = 1 - \frac{Z^s_{inter}\cdot Z^{aug}_{inter}}{\|Z^s_{inter}\|\cdot\|Z^{aug}_{inter}\|} \tag{19}$$

where the $\mathcal{L}^{inter}_{con}$ is computed by the cosine similarity. The $\mathcal{L}^{inter}_{con}$ is minimized to reduce the difference between the Z^s_{inter} and Z^{aug}_{inter} for learning domain-invariant inter-class relations. In the end, the Z^s_{inter} and Z^{aug}_{inter} are concatenated as the output Z_{inter} of the DIERL.

3.3. The intra-class relations learning

The intra-class relations learning (ITARL) aims to learn intra-class relations which are beneficial for improving class discriminability. Since there is no obvious intra-class relation in stuff classes (including road, sidewalk, building, wall, fence, vegetation, terrain, and sky), only the intra-class relation of instance classes (including pole, light, sign,

[1] https://github.com/huawei-noah/Efficient-AI-Backbones/tree/master/vig_pytorch.

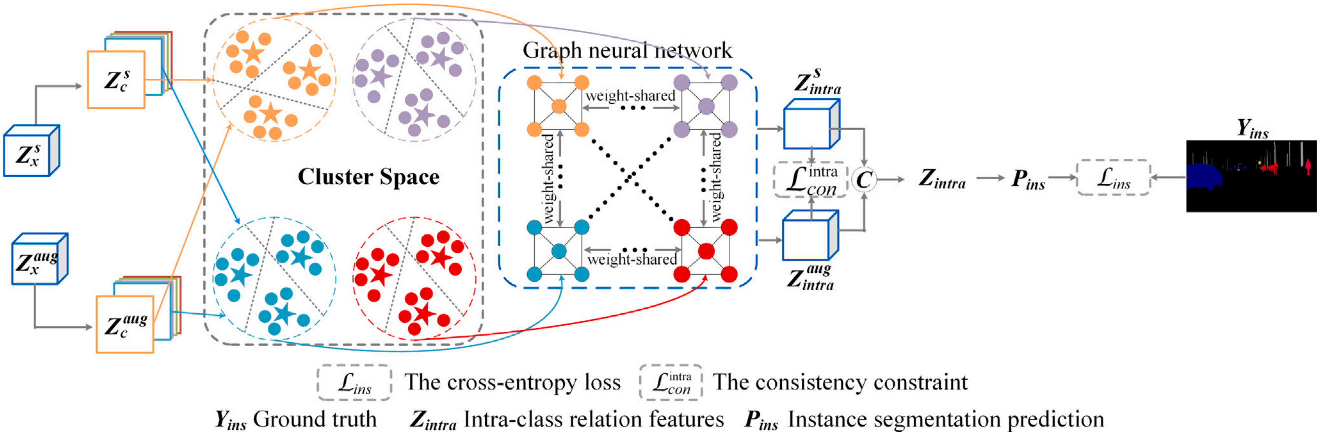


Fig. 3. The architecture of the ITARL. In cluster space, each class-specific feature is divided into N clusters using the K-means cluster method. Each cluster indicates the different parts of instance classes, such as the head, arm, leg, and body of the "person" class, where the pentagram denotes the centroid of the cluster and the nearest circle to the centroid represents the feature that may belong to the same part of instance classes. Similar to the DIERL, the feature of each part in an instance class is viewed as a set of unordered nodes of the graph, and the edges are added for all nodes to build the graph $\mathcal{G}_i = \{\mathcal{V}_i, \mathcal{E}_i\}$ which is put into a weight-shared GNN to learn intra-class relations, where $i \in [1, C]$ and C is the number of the instance classes.

Algorithm 1 GVRL algorithm.

```
Input: The source domain images, the augmented domain images, and labels $(X^s, X^{aug}, Y)$
Output: The trained model $F$
Initialize network parameters $\theta$.
$iter \leftarrow 1$
$max\_iter \leftarrow STOP\_NUM$
while $iter \le max\_iter$ do
    $Z_x \leftarrow F(X^s, X^{aug}; \theta)$
    $Z_x^s, Z_x^{aug} \leftarrow$ split$(Z_x)$
    $Z_c^s \leftarrow$ Eq. (1)
    $Z_c^{aug} \leftarrow$ Eq. (2)
    Initialize the prototypes $Z_p$ by Eq. (3)
    Update the prototypes $Z_p$ by Eq. (4)
    Compute $\mathcal{L}_{mpcl}$ with Eq. (5)
    $Z_{inter} \leftarrow$ feed $Z_c^s$ and $Z_c^{aug}$ into the DIERL
    Compute $\mathcal{L}_{con}^{inter}$ with Eq. (19)
    $Z_{intra} \leftarrow$ feed $Z_c^s$ and $Z_c^{aug}$ into the ITARL
    Compute $\mathcal{L}_{con}^{intra}$ with Eq. (20)
    Compute $\mathcal{L}_{ins}$ with Eq. (21)
    $Z_f \leftarrow$ concat $(Z_x, Z_{inter}, Z_{intra})$
    $P_f \leftarrow$ softmax(upsampling $(Z_f)$)
    $P_f^s, P_f^{aug} \leftarrow$ split$(P_f)$
    Compute $\mathcal{L}_{con}^p$ with Eq. (7)
    Compute $\mathcal{L}_{seg}$ with Eq. (6)
    Train $\mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{mpcl} + \lambda_2 \mathcal{L}_{ins} + \lambda_3 (\mathcal{L}_{con}^{inter} + \mathcal{L}_{con}^{intra} + \mathcal{L}_{con}^p)$
    Backward
    $iter \leftarrow iter + 1$
end while
```

person, rider, car, truck, bus, train, motorcycle, and bicycle) is considered. The architecture of the ITARL is shown in Fig. 3. Specifically, for each class-specific feature $z_i \in Z_c$, the z_i is divided into N clusters by using the K-means cluster method, where each cluster indicates the different parts of instance classes, such as the head, arm, leg, and body of the "person" class. In particular, N is a hyper-parameter since it is unknown how many parts each class is divided into. The feature of each part in an instance class is viewed as a set of unordered nodes of the graph, and the edges are added for all nodes to build the graph $\mathcal{G}_i = \{\mathcal{V}_i, \mathcal{E}_i\}$, where $i \in [1, C]$ and C is the number of the instance classes. Similar to the DIERL, the graph $\mathcal{G}_i$ is put into a weight-shared GNN (Han et al., 2022) to learn intra-class relations Z_{intra} for each instance class. To ensure the intra-class relations Z_{intra} are consistent across domains, the consistency constraint between the Z_{intra}^s and Z_{intra}^{aug} is leveraged, which is denoted as follows:

$$\mathcal{L}_{con}^{intra}(Z_{intra}^s, Z_{intra}^{aug}) = 1 - \frac{Z_{intra}^s \cdot Z_{intra}^{aug}}{\|Z_{intra}^s\| \cdot \|Z_{intra}^{aug}\|} \tag{20}$$

where the $\mathcal{L}_{con}^{intra}$ is computed by the cosine similarity. The $\mathcal{L}_{con}^{intra}$ is minimized to reduce the difference between the Z_{intra}^s and Z_{intra}^{aug} for learning domain-invariant intra-class relations. Finally, the Z_{intra}^s and Z_{intra}^{aug} are concatenated as the output Z_{intra} of the ITARL. Moreover, to ensure the Z_{intra} is effective in semantic segmentation, the Z_{intra} is utilized to predict the instance segmentation map P_{ins} and the Z_{intra} is optimized by the cross-entropy loss $\mathcal{L}_{seg}^{ins}$ between the P_{ins} and the corresponding ground-truth label Y_{ins}, which is denoted as follows:

$$\mathcal{L}_{seg}^{ins}(P_{ins}, Y_{ins}) = -Y_{ins} \cdot \log(P_{ins}) \tag{21}$$

where P_{ins} is obtained by upsampling the intra-class relations Z_{intra}. The ground-truth label Y_{ins} is generated by using an instance classes mask M_{ins}, which is denoted as follows:

$$Y_{ins} = Y_s \circ M_{ins} \tag{22}$$

where $\circ$ is denoted the Hadamard product of the matrix.

3.4. The training loss

The overall objective of our approach can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{mpcl} + \lambda_2 \mathcal{L}_{ins} + \lambda_3 (\mathcal{L}_{con}^{inter} + \mathcal{L}_{con}^{intra} + \mathcal{L}_{con}^p) \tag{23}$$

where λ_i is weights coefficients. The training process is summarized in Algorithm 1.

4. Experiments

4.1. Datasets

Five standard single-source domain generalization semantic segmentation benchmarks are utilized to evaluated the proposed approach, including "G→{C, B, M}", "S→{C, M, B}", and "C→{B, G, S}". The "G", "S", "C", "M", and "B" respectively denote the GTA5 (Richter, Vineet, Roth, & Koltun, 2016), SYNTHIA-RAND-CITYSCAPES (abbreviated as **SYNTHIA (S)**) (Ros, Sellart, Materzynska, Vazquez, & Lopez, 2016), Cityscapes (Cordts et al., 2016), Mapillary (Neuhold, Ollmann, Rota Bulo, & Kontschieder, 2017), and BDD100K (Yu et al., 2020) datasets. In particular, the form of describing benchmark, e.g., "G→{C,

Table 1
Comparison between DGSS methods on the “G→{C, B, M}” benchmark based on the ResNet-101, ResNet-50, ShuffleNetV2, and MobileNetV2 backbones. The best results are marked in bold.

Backbone	Methods	Venue	C	B	M	Avg
ResNet-50	SAN-SAW (Peng et al., 2022)	CVPR’22	39.8	37.3	41.9	39.7
	SHADE (Zhao et al., 2022)	ECCV’22	44.7	39.3	43.3	42.4
	SPC (Huang et al., 2023)	CVPR’23	44.1	40.5	45.5	43.4
	ATF (Li, Li, Wang, Ren, & Guo, 2023)	AAAI’23	39.9	38.8	42.8	40.5
	DPCL (Yang et al., 2023)	AAAI’23	44.9	40.2	46.7	43.9
	CDPCL (Liao et al., 2023a)	ACM MM’23	42.2	39.5	42.4	41.4
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP’23	41.5	37.8	43.4	40.9
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT’23	44.5	36.5	44.3	41.8
	Our	–	**45.5**	**42.1**	**46.9**	**44.8**
ResNet-101	SAN-SAW (Peng et al., 2022)	CVPR’22	45.3	41.2	40.8	42.4
	SHADE (Zhao et al., 2022)	ECCV’22	46.7	43.7	45.5	45.3
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP’23	48.0	40.3	47.4	45.2
	Our	–	**48.5**	**45.1**	**48.7**	**47.4**
ShuffleNetV2	CDPCL (Liao et al., 2023a)	ACM MM’23	35.4	35.9	36.3	35.9
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP’23	34.5	33.4	36.6	34.8
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT’23	35.3	33.1	37.6	35.3
	Our	–	**36.9**	**36.4**	**39.2**	**37.5**
MobileNetV2	CDPCL (Liao et al., 2023a)	ACM MM’23	36.9	33.7	36.7	35.8
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP’23	36.3	33.0	36.2	35.2
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT’23	35.8	33.2	36.0	35.0
	Our	–	**37.2**	**36.8**	**38.5**	**37.5**

B, M}”, means the model is trained on the GTAV (G) dataset and tested the model on the Cityscapes (C), BDD100K (B), and Mapillary (M) datasets. The GTA5 (Richter et al., 2016) and SYNTHIA (Ros et al., 2016) are two synthetic datasets, which respectively contains 24 966 synthetic images with a resolution of 1914 × 1052 and 9400 photo-realistic images with a resolution of 1280 × 760. The Cityscapes (Cordts et al., 2016), Mapillary (Neuhold et al., 2017), and BDD100K (Yu et al., 2020) are three real-world datasets, which respectively contains 2975 real-world images with a resolution of 2048 × 1024, 20 000 real-world images with different resolutions, and 8000 real-world images with a resolution of 1280 × 720. We split each dataset followed by previous works (Choi et al., 2021; Kim et al., 2022; Peng et al., 2022; Zhao et al., 2022).

4.2. Implementation details

Model. The DeepLabV3+ (Chen, Zhu, Papandreou, Schroff, & Adam, 2018) with the ResNet-101 (He, Zhang, Ren, & Sun, 2016), ResNet-50 (He et al., 2016), ShuffleNetV2 (Ma, Zhang, Zheng, & Sun, 2018), and MobileNetV2 (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) backbones are respectively used as the segmentation network to conduct experiments.

Training details. The experiments are conducted on two GTX 3090 GPUs. The maximum number of iterations is set to 40k steps. The input images are randomly cropped to the resolution of 768 × 768. The weights coefficient λ_i shown in Eq. (23) is set as: $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$. The trade-off parameter m shown in Eq. (4) is set as 0.9. The temperature parameter τ shown in Eq. (5) is set as 0.8. The hyper-parameter N of the intra-class relation learning module is set as 4.

Evaluation metric. The mean Intersection-Over-Union value ($mIoU = \frac{TP}{TP+FP+FN}$) is utilized as the evaluation metric in experiments, where TP, FP, and FN are denoted as the predicted results of true positive, false positive, and false negative respectively.

4.3. Comparison with DG methods

Following Peng et al. (2022) and Choi et al. (2021), for the ResNet-101 (He et al., 2016), ResNet-50 (He et al., 2016), ShuffleNetV2 (Ma et al., 2018), and MobileNetV2 (Sandler et al., 2018) backbones, our approach is evaluated on five standard benchmarks, including “G→{C, B, M}”, “S→{C, M, B}”, and “C→{B, G, S}”.

Several recent approaches (Ding, Xue, Xia, Schiele, & Dai, 2023; Huang et al., 2023; Li et al., 2023; Liao et al., 2023a; Peng et al., 2022; Yang et al., 2023; Zhang, Tian, Liao, Hua et al., 2023; Zhang, Tian, Liao, Zhang et al., 2023; Zhao et al., 2022) are compared with the proposed GVRL. For the four backbones, the results are shown in Tables 1, 2, and 3. Our approach achieves superior performance than recent methods in three generalization settings.

The “G→{C, B, M}” task. As shown in Table 1, for the ResNet-50 backbone, compared with the state-of-the-art (SOTA) method SPC (Huang et al., 2023) and DPCL (Yang et al., 2023), our approach achieves a significant average improvement of 1.4% and 0.9% in terms of mIoU. For the ResNet-101 backbone, compared with the SOTA method SIL (Zhang, Tian, Liao, Hua et al., 2023), the proposed GVRL achieves a significant average improvement of 2.2% in terms of mIoU. For the ShuffleNetV2 backbone, compared with the SOTA method CDPCL (Liao et al., 2023a), the proposed GVRL achieves a significant average improvement of 1.6% in terms of mIoU. For the MobileNetV2 backbone, compared with the SOTA method CDPCL (Liao et al., 2023a), the proposed GVRL achieves a significant average improvement of 1.7% in terms of mIoU.

The “S→{C, M, B}” task. As shown in Table 2, for the ResNet-50 backbone, compared with the state-of-the-art (SOTA) method ATF (Li et al., 2023) and CDPCL (Liao et al., 2023a), the proposed approach achieves a significant average improvement of 1.1% and 1.3% in terms of mIoU. For the ResNet-101 backbone, compared with the SOTA method ATF (Li et al., 2023), the proposed GVRL achieves a significant average improvement of 1.7% in terms of mIoU. For the ShuffleNetV2 and MobileNetV2 backbones, compared with the SOTA method CDPCL (Liao et al., 2023a), the proposed GVRL achieves significant average improvement of 4.0% and 5.2% in terms of mIoU, respectively.

The “C→{B, G, S}” task. As shown in Table 3, for the ResNet-50 backbone, compared with the state-of-the-art (SOTA) method HGFormer (Ding et al., 2023), our approach achieves a significant average improvement of 1.6% in terms of mIoU. For the ResNet-101 backbone, compared with the recent method SAN-SAW (Peng et al., 2022) and SIL (Zhang, Tian, Liao, Hua et al., 2023), the proposed GVRL achieves significant average improvement of 2.6% and 3.8% in terms of mIoU, respectively. For the ShuffleNetV2 and MobileNetV2 backbones, compared with the SOTA method CDPCL (Liao et al., 2023a), the proposed GVRL achieves significant average improvement of 1.4% in terms of mIoU, respectively.

Table 2
Comparison between DGSS methods on the "S→{C, B, M}" benchmark based on the ResNet-101, ResNet-50, ShuffleNetV2, and MobileNetV2 backbones. The best results are marked in bold.

Backbone	Methods	Venue	C	B	M	Avg
ResNet-50	SAN-SAW (Peng et al., 2022)	CVPR'22	38.9	35.2	34.5	36.2
	ATF (Li et al., 2023)	AAAI'23	39.5	36.1	37.1	37.6
	CDPCL (Liao et al., 2023a)	ACM MM'23	41.2	35.4	35.5	37.4
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP'23	39.2	30.4	31.0	33.5
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT'23	39.2	29.8	33.5	34.2
	Our	–	**41.8**	**36.9**	**37.4**	**38.7**
ResNet-101	SAN-SAW (Peng et al., 2022)	CVPR'22	40.9	36.0	37.3	38.1
	ATF (Li et al., 2023)	AAAI'23	41.3	36.9	39.1	39.1
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP'23	40.9	33.4	33.4	35.9
	Our	–	**42.2**	**37.5**	**39.6**	**39.8**
ShuffleNetV2	CDPCL (Liao et al., 2023a)	ACM MM'23	35.4	24.7	28.0	29.4
	Our	–	**36.5**	**28.8**	**34.8**	**33.4**
MobileNetV2	CDPCL (Liao et al., 2023a)	ACM MM'23	35.5	23.9	28.4	29.3
	Our	–	**38.2**	**29.8**	**35.5**	**34.5**

Table 3
Comparison between DGSS methods on the "C→{B, G, S}" benchmark based on the ResNet-101, ResNet-50, ShuffleNetV2, and MobileNetV2 backbones. The best results are marked in bold.

Backbone	Methods	Venue	B	G	S	Avg
ResNet-50	SAN-SAW (Peng et al., 2022)	CVPR'22	53.0	47.3	28.3	42.9
	SHADE (Zhao et al., 2022)	ECCV'22	51.0	48.6	27.6	42.4
	DPCL (Yang et al., 2023)	AAAI'23	52.3	46.0	26.6	41.6
	HGFormer (Ding et al., 2023)	CVPR'23	51.5	50.4	30.1	44.0
	CDPCL (Liao et al., 2023a)	ACM MM'23	52.2	46.8	31.9	43.6
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP'23	50.8	44.8	30.9	42.2
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT'23	52.0	45.1	30.0	42.4
	Our	–	**53.1**	**51.1**	**32.5**	**45.6**
ResNet-101	SAN-SAW (Peng et al., 2022)	CVPR'22	54.7	48.8	30.2	44.6
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP'23	53.0	45.0	32.3	43.4
	Our	–	**56.1**	**50.5**	**34.9**	**47.2**
ShuffleNetV2	CDPCL (Liao et al., 2023a)	ACM MM'23	43.6	41.1	28.1	37.6
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP'23	43.9	40.6	26.1	36.9
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT'23	44.1	40.3	27.5	37.3
	Our	–	**45.9**	**41.7**	**29.4**	**39.0**
MobileNetV2	CDPCL (Liao et al., 2023a)	ACM MM'23	48.4	43.2	27.3	39.6
	SIL (Zhang, Tian, Liao, Hua et al., 2023)	TIP'23	46.3	42.2	27.2	38.6
	FGSS (Zhang, Tian, Liao, Zhang et al., 2023)	TCSVT'23	46.6	42.3	27.6	38.8
	Our	–	**49.2**	**45.0**	**28.9**	**41.0**

Table 4
Ablation experiments of the proposed DIERL and ITARL in the "G→{C, B, M}" generalization setting based on the ShuffleNetV2 (Ma et al., 2018) backbone. The best results are marked in bold.

Methods	Group ID	DIERL	ITARL	Train on GTA5 (G)			
				C	B	M	Avg
Baseline	A	–	–	31.0	32.1	35.3	30.7
Ours	B	√	–	36.2	35.9	38.5	36.9
	C		√	35.4	35.1	38.3	36.3
	D	√	√	**36.9**	**36.4**	**39.2**	**37.5**

4.4. Ablation study

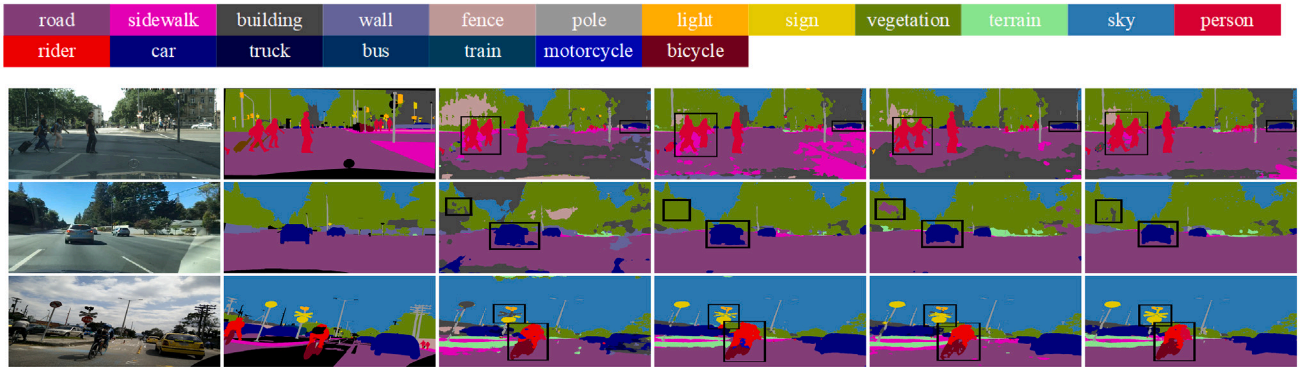
In this section, the effectiveness of each component in the proposed GVRL is verified from two aspects on the "G→{C, B, M}" task based on the ShuffleNetV2 backbone. First, the ablation studies of the proposed DIERL and ITARL are conducted to verify their effectiveness. Second, the ablation studies of L^{p}_{con}, L^{inter}_{con}, L^{intra}_{con}, L_{ins}, and L_{mpcl} are conducted to verify the effectiveness of each loss function.

The ablation studies of the proposed DIERL and ITARL. The results are shown in Table 4. Compared with the baseline (*cf.* group A), the average performance on three datasets after using the DIERL (*cf.* group B) or using the ITARL (*cf.* group C) or simultaneously using these two modules (*cf.* group D) achieves significant improvement. The improved performance demonstrates that the learned domain-invariant inter-class relations or the learned intra-class relations are beneficial for domain generalization semantic segmentation. Compared with the performance of groups A, B, and C, group D (our approach) achieves superior performance. This result demonstrates that the DIERL and ITARL are complementary. Furthermore, the performances of per-class IoU, mIoU, mIoU of instance classes, and mIoU of stuff classes on the "GTA5→Cityscapes" task are shown in Table 5 to further analyze the effectiveness of our proposed modules. Specifically, first, compared with baseline (*cf.* group A), the performance of group B which uses the DIERL respectively achieves the improvement on stuff and instance classes which demonstrates that the DIERL can learn domain-invariant inter-class relations for improving the generalization ability. Second, the comparison between the baseline (*cf.* group A) and it with the ITARL (*cf.* group C) is analyzed. From Table 5, after using the ITARL, the performance of stuff and instance classes are also improved. In particular, the performance of instance classes achieves a significant improvement of 5.9% in terms of mIoU. It demonstrates that the ITARL can learn more discriminative features of instance classes to improve class discriminability. Finally, the comparison between groups A, B, C, and D is analyzed. The overall performance of group D achieves the improvement of 5.9%, 0.7%, and 1.5% in terms of mIoU compared with groups A, B, and C. It demonstrates that simultaneously using the DIERL and ITARL can improve the generalization ability and class discriminability.

Moreover, the visualization of the ablation study is shown in Fig. 4, where the black box marks the improvement of using the DIERL and

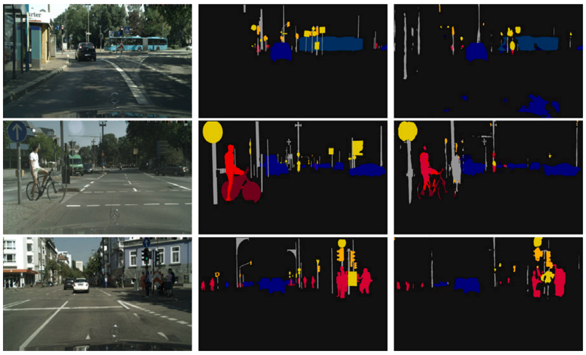
Table 5
Quantitative comparison results on Cityscapes when adapted from GTA5 in terms of per-class IoUs and mIoU (%). The best results are marked in bold and the second best results are underlined. The italic indicates instance classes.

Backbones	Methods	mIoU	Ins mIoU	Stuf mIoU	GTA5 to Cityscapes per-class IoU (%)																		
					road	sidewalk	building	wall	fence	*pole*	*light*	*sign*	vegetation	terrain	sky	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motorcycle*	*bicycle*
ShuffleNetV2 (Ma et al., 2018)	Group A (Baseline)	31.0	17.5	49.4	78.9	30.2	73.6	17.1	17.4	*20.5*	*11.9*	*7.7*	80.4	33.7	64.2	*39.7*	*10.5*	*71.7*	*15.6*	*10.4*	*0.0*	*3.7*	*0.8*
	Group B (DIERL)	36.2	21.4	**56.6**	**87.3**	35.8	**79.8**	**29.8**	**24.7**	*23.8*	*26.9*	*15.6*	**83.6**	**39.2**	**72.5**	*43.6*	*4.8*	*76.1*	*16.1*	*11.5*	*2.7*	*10.4*	*3.3*
	Group C (ITARL)	35.4	23.5	51.7	79.8	36.2	73.2	22.8	18.3	*24.7*	*29.5*	*18.6*	80.7	36.6	66.2	*48.5*	*5.7*	*79.5*	*21.9*	*8.8*	*1.4*	*12.7*	*6.6*
	Group D (Ours)	**36.9**	**24.4**	54.2	84.5	**37.4**	76.0	26.0	20.3	*27.9*	*29.8*	*19.0*	**83.6**	37.6	68.3	*47.2*	*6.0*	*81.1*	*21.6*	*13.3*	*2.1*	*15.0*	*5.2*



(a) Input image (b) Ground truth (c) Group A (Baseline) (d) Group B (DIERL) (e) Group C (ITARL) (f) Group D (Ours)

Fig. 4. Visualization of ablation study (*cf.* Table 4) in the G→{C, B, M} task. The visualization results from three datasets (including Cityscapes, BDD, Mapillary) are respectively shown in three rows.



(a) RGB (b) GT (c) Prediction

Fig. 5. Segmentation results visualization of instance classes.

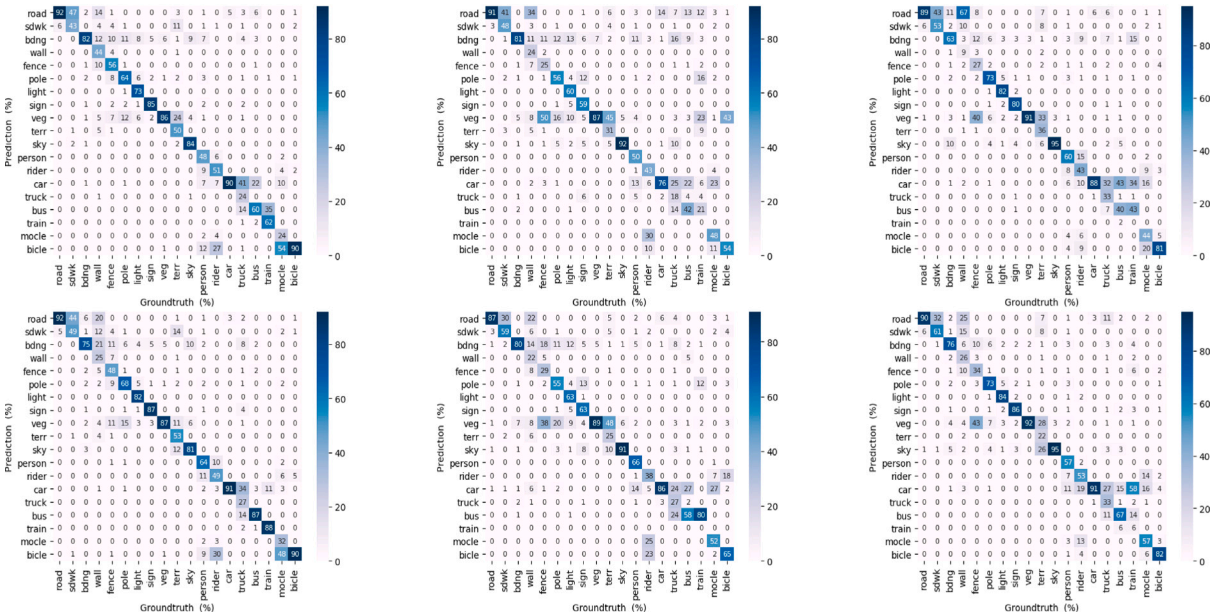
ITARL modules. For one thing, compared with group A, group B which uses the DIERL achieves more accurate segmentation results overall. For example, in the black box area of group A shown in the second and last rows of Fig. 4(c), the vegetation and sidewalk are inaccurately segmented to the terrain and road which is an anti-realism prediction in the real world. After using the DIERL, group B achieves more accurate segmentation results of the vegetation and sidewalk which is shown in the black box area in the second and last row of Fig. 4(d). It demonstrates that the learning of the domain-invariant inter-class relation, which provide prior knowledge about the real world, can alleviate these anti-realism predictions. For another, compared with group A, group C which uses the ITARL achieve better segmentation result of instance classes. For example, the comparison between the black box area of groups A and C which are shown in fours rows of Fig. 4(c) and (e), group C obtains more accurate segmentation results of the classes of the person, car, bicycle, sign, and rider. Finally, jointly using the DIERL and ITARL modules, our approach (*cf.* group D) on all classes achieves more accurate segmentation results overall than groups A, B, and C, which are shown in the black boxes of Fig. 4(f).

Notably, as shown in Fig. 4(e) and (f), on one hand, the segmentation results of Group C are better than Group D in some instance classes, e.g., the "person" and the "rider" classes shown in the first and last row. These phenomena may be caused by that Group C mainly focuses on learning intra-class relations of instance classes to improve class discriminability, which achieves accurate segmentation results on instance classes. While Group D needs to learn domain-invariant inter-class relations and intra-class relations of instance classes to improve the generalization ability and class discriminability simultaneously, which may affect the improvement of class discriminability on instance classes compared with Group C. On the other hand, the segmentation results of Group D are overall more accurate than Group C, e.g., in the "road", "vegetation", "wall", and the "sign" classes. Thus, although the segmentation results of some instance classes may be suboptimal compared to Group C, the overall segmentation results of Group D are more accurate than Group C. In addition, the segmentation results of instance classes are visualized to verify the effectiveness of the ITARL. As shown in Fig. 5(b) and (c), the visual samples between the ground truth and the prediction are compared. From Fig. 5(c), the instance classes are successfully segmented which demonstrates the intra-class relations of instance classes are effectively learned by the ITARL.

The ablation studies of L^{p}_{con}, L^{inter}_{con}, L^{intra}_{con}, L_{ins}, and L_{mpcl}. The results are shown in Table 6. Compared with the baseline (*cf.* group A), the average performance on three datasets significantly improves by progressively adding each loss function. Specifically, compared with the baseline (*cf.* group A), to keep the consistent prediction between the source and augmented domains brings a 2.3% improvement in terms of mIoU. The use of the L^{inter}_{con} and L^{intra}_{con} bring the 0.8% and 0.9% improvements in terms of mIoU, respectively. Moreover, After using the L_{mpcl} to learn domain-invariant class-specific features, the performance achieves an improvement of 1.1% in terms of mIoU. It demonstrates that learning domain-invariant class-specific features probably mines

Table 6
Ablation experiments of L_{con}^{p}, L_{con}^{inter}, L_{con}^{intra}, L_{ins}, and L_{mpcl} in the "G→{C, B, M}" generalization setting based on the ShuffleNetV2 (Ma et al., 2018) backbone. The best results are marked in bold.

Backbone	Methods	Group ID	L_{con}^{p}	DIERL	ITARL		L_{mpcl}	Train on GTA5 (G)			
				L_{con}^{inter}	L_{con}^{intra}	L_{ins}		C	B	M	Avg
ShuffleNetV2	Baseline	A	–	–	–	–	–	31.0	32.1	35.3	30.7
	Ours	B	√	–	–	–	–	32.7	33.4	35.9	34.0
		C	√	√	–	–	–	34.1	33.9	36.3	34.8
		D	√	√	√	–	–	34.8	35.2	37.1	35.7
		E	√	√	√	√	–	35.3	35.8	38.0	36.4
		F	√	√	√	√	√	**36.9**	**36.4**	**39.2**	**37.5**



(a) Cityscapes (b) BDD (c) Mapillary

Fig. 6. Visualization of confusion matrix of the group B and group D (*cf.* Table 4) experiments in the G→{C, B, M} setting.

more stable inter-class relations which is beneficial for improving the performance in the unseen domains. In particular, the use of L_{ins}, which aims to learn effective features for segmenting instance classes, brings a 0.7% improvement in terms of mIoU.

4.5. Confusion matrix analysis

The confusion matrices of group B (*cf.* Table 4) and group D (our approach) is provided to analyze the class confusion problem. The confusion matrices are shown in Fig. 6. The misclassification of instance classes is mainly analyzed. Compared with group B, the class confusion problem of instance classes is alleviated by our approach on four out-of-distribution datasets, where the misclassification of "light", "sign", "car", "truck", "bus", "train", and "motorcycle" is significantly decreased. It demonstrates that the learning of intra-class relations is beneficial for improving class discriminability.

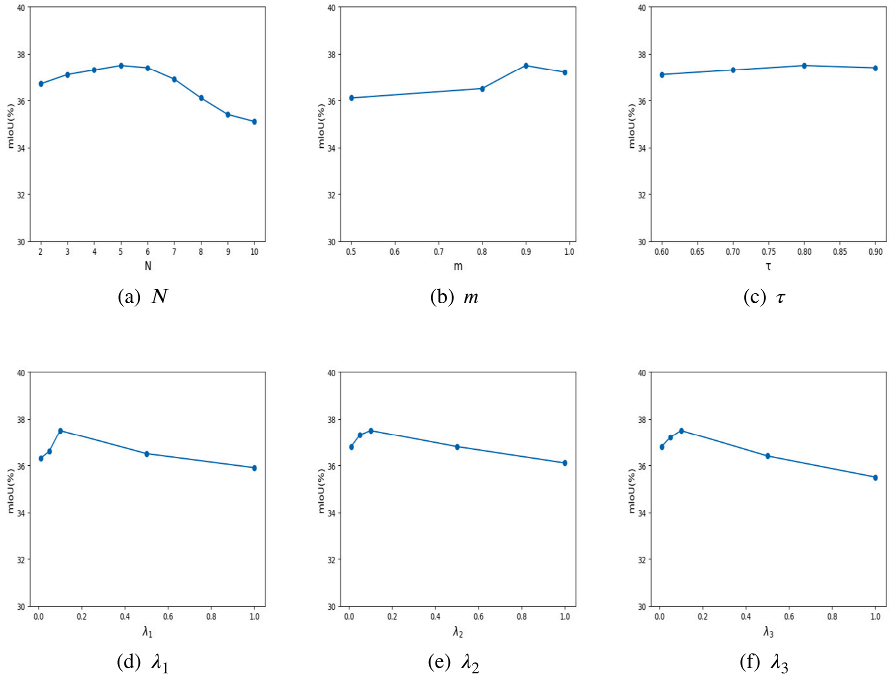
Specifically, the average misclassification of "pole", "light", "sign", "person", "rider", "car", "truck", "bus", "train", "motorcycle", and "bicycle" is respectively decreased 1.8%, 3.3%, 4.5%, 7.0%, 5.8%, 5.0%, 3.0%, 20.5%, 8.0%, 10.0%, and 2.3%. Overall, the degradation in misclassification of instance classes demonstrates the effectiveness of our approach in alleviating the class confusion problem.

4.6. Performance dependency of hyper-parameters

We further investigate the performance dependency of the hyper-parameters N, m, τ, λ_1, λ_2, and λ_3. For these hyper-parameters, we use a grid search over $N \in \{2,3,4,5,6,7,8,9,10\}$, $m \in \{0.5, 0.8, 0.9, 0.99\}$, $\tau \in \{0.6, 0.7, 0.8, 0.9\}$, and λ_1, λ_2, $\lambda_3 \in \{1.0, 0.5, 0.1, 0.05, 0.01\}$. The results are shown in Fig. 7.

The hyper-parameter N affects the learning of intra-class relations since N is utilized to determine how many parts to divide an instance class into. The smaller value of N means a coarse division for an instance class, such as the "person" class is divided into head and body. The higher value of N means more fine-grained division for an instance class, such as the "person" class is divided into head, neck, body, left hand, right hand, left leg, and right leg. Intuitively, the coarse division may weaken to capture the intra-class variations and the fine-grained division may increase the sensitivity of the model to detail differences that contain distribution shifts, e.g., style, color, etc (Sohoni, Dunnmon, Angus, Gu, & Ré, 2020). Moreover, more fine-grained division increases the computational cost of the clustering. Thus, we conduct the hyper-parameter experiments to select a suitable value for N. From experimental results shown in Fig. 7(a), the performance first increases and then decreases with the increasing of N, where the value of N is set to 5 to achieve the best performance 37.7% in terms of mIoU. Considering the balance of the performance and computational cost, the value of N is set to 4, achieving a performance of 37.5% in terms of mIoU. This performance only has a minor difference compared with the performance of $N = 5$, while the computational cost of $N = 4$ is less than $N = 5$.

The hyper-parameter m is used to update the prototype Z_p and the temperature parameter τ is used to affect the sensitivity to the similarity between positive and negative samples. In particular, the smaller value of τ may lead to a higher difference between the positive samples and the most similar negative samples, which causes the model



(a) N (b) m (c) τ

(d) λ_1 (e) λ_2 (f) λ_3

Fig. 7. Quantitative comparisons for the hyperparameters, including N, m, τ, λ_1, λ_2, and λ_3. All experiments are based on the ShuffleNetV2 (Ma et al., 2018) in the G→{C, B, M} setting. We show the average performance of three datasets for all cases.

Table 7
Computational cost analysis. Params denotes parameters. FLOPs denotes floating point operations.

Method	FLOPs (G)	Params (M)	Inference time (ms)
Baseline (Chen et al., 2018)	155.92	45.07	131.22
DPCL (Yang et al., 2023)	594.32	56.46	441.89
Ours	221.80	60.26	375.00

to be more inclined to separate these samples. On the contrary, the higher value of τ may lead to reduce the difference between positive and negative samples. From the experimental results shown in Fig. 7(b) and (c), m and τ are set to the values of 0.9 and 0.8 to achieve the best performance.

The hyper-parameters λ_1, λ_2, and λ_3 are utilized to balance the contributions between the cross-entropy loss of semantic segmentation, the learning of domain-invariant class-specific features, and the learning of inter-class and intra-class relations. From the experimental results shown in Fig. 7(d), (e), and (f), λ_1, λ_2, and λ_3 are set to the weights as 0.1, 0.1, and 0.1, which achieves the best performance.

We fix hyperparameters and train our model on all settings.

4.7. Model complexity analysis

The computational cost of the proposed approach is compared with the baseline (Chen et al., 2018) and the DPCL (Yang et al., 2023). The comparison experiments are conducted on a single NVIDIA RTX 3090 GPU. The comparison methods and the proposed approach are based on DeepLabV3+ with the ResNet-50 backbone. The input size of an image is set as 1024 × 512 × 3. From Table 7, although the computational cost of the proposed approach is more than the baseline (Chen et al., 2018), there is less computational cost of the proposed approach than the recent method DPCL (Yang et al., 2023), including floating point operations and inference time.

4.8. Limitations and future works

Although our approach achieves a significant improvement over baseline and other methods in the average performance and the performance of instance classes, the proposed approach still has some limitations. First, the performance improvement of stuff classes is unsatisfactory since our approach ignores mining the intra-class relations of stuff classes due to there being no obvious intra-class relations within stuff classes. Second, the tiny pyramid graph neural network is used for the learning of inter-class and intra-class relations, which increases the computational cost of the proposed approach. Thus, in future work, we will explore how to improve the performance of stuff classes (e.g., road, sky, and building) and design a parameter-efficient visual relation learning method to reduce the computational cost.

5. Conclusion

In this paper, a novel approach, named generalized visual relations learning (GVRL), is proposed to learn domain-invariant inter-class relations and intra-class relations for domain generalization semantic segmentation. Specifically, the GVRL contains a domain-invariant inter-class relation learning (DIERL) and an intra-class relation learning (ITARL). For one thing, the DIERL aims to learn the domain-invariant inter-class relations by constructing a prototype of each class and using the consistency constraint of the inter-class relations between the source and augmented domains for improving the generalization ability. For another, the ITARL aims to learn intra-class relations of instance classes (e.g., person, car, and rider) for improving class discriminability. Extensive experiments demonstrate that our approach achieves superior performance over current approaches on domain generalization segmentation tasks.

CRediT authorship contribution statement

Zijun Li: Designed the initial framework of the proposed method, Discussed and optimized the design, Planned the experiments, Analyzed the results, Writing – original draft, Revised the introduction, Proposed method description, Final version of the manuscript. **Muxin Liao:** Discussed and optimized the design, Planned the experiments, Analyzed the results, Revised the paper, Provided critical feedback, Carried out the experiments and assisted with the analysis of the results, Revised the abstract, Introduction, Statement of the proposed framework, Illustrate the motivation of the proposed generalized visual relations learning approach, Revised the introduction, Gave some helpful comments on the analysis of the experimental results, Final version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Asutkar, S., Chalke, C., Shivgan, K., & Tallur, S. (2023). TinyML-enabled edge implementation of transfer learning framework for domain generalization in machine fault diagnosis. *Expert Systems with Applications*, *213*, Article 119016.

Baradel, F., Neverova, N., Wolf, C., Mille, J., & Mori, G. (2018). Object level visual reasoning in videos. In *Proceedings of the European conference on computer vision* (pp. 105–121).

Barbato, M. P., Piccoli, F., & Napoletano, P. (2024). Ticino: A multi-modal remote sensing dataset for semantic segmentation. *Expert Systems with Applications*, Article 123600.

Borse, S., Park, H., Cai, H., Das, D., Garrepalli, R., & Porikli, F. (2022). Panoptic, instance and semantic relations: A relational context encoder to enhance panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1269–1279).

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., & Kalantidis, Y. (2019). Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 433–442).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision* (pp. 801–818).

Choi, S., Jung, S., Yun, H., Kim, J. T., Kim, S., & Choo, J. (2021). 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* (pp. 11580–11590).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Ding, X., Shen, C., Zeng, T., & Peng, Y. (2022). SAB Net: A semantic attention boosting framework for semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Ding, J., Xue, N., Xia, G.-S., Schiele, B., & Dai, D. (2023). Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15413–15423).

Dong, Z., Niu, S., Gao, X., Li, J., & Shao, X. (2024). Uncertainty-weighted prototype active learning in domain adaptive semantic segmentation. *Expert Systems with Applications*, *245*, Article 123094.

Elhassan, M. A., Huang, C., Yang, C., & Munea, T. L. (2021). DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes. *Expert Systems with Applications*, *183*, Article 115090.

Fan, C., Chen, W., Tian, J., Li, Y., He, H., & Jin, Y. (2024). Unlock the potential of counterfactually-augmented data in out-of-distribution generalization. *Expert Systems with Applications*, *238*, Article 122066.

Fan, X., Zhou, W., Qian, X., & Yan, W. (2024). Progressive adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images. *Expert Systems with Applications*, *238*, Article 121999.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3146–3154).

González-Collazo, S. M., Balado, J., González, E., & Nurunnabi, A. (2023). A discordance analysis in manual labelling of urban mobile laser scanning data used for deep learning based semantic segmentation. *Expert Systems with Applications*, *230*, Article 120672.

Guo, C., Zhao, Z., Ren, J., Wang, S., Liu, Y., & Chen, X. (2024). Causal explaining guided domain generalization for rotating machinery intelligent fault diagnosis. *Expert Systems with Applications*, *243*, Article 122806.

Han, K., Wang, Y., Guo, J., Tang, Y., & Wu, E. (2022). Vision GNN: An image is worth graph of nodes. *Advances in Neural Information Processing Systems*.

He, X., Liu, J., Wang, W., & Lu, H. (2022). An efficient sampling-based attention network for semantic segmentation. *IEEE Transactions on Image Processing*, *31*, 2850–2863.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Hong, H.-S., Kumar, A., & Lee, D.-G. (2024). Robust unsupervised domain adaptation by retaining confident entropy via edge concatenation. *Expert Systems with Applications*, *238*, Article 122120.

Hu, H., Ji, D., Gan, W., Bai, S., Wu, W., & Yan, J. (2020). Class-wise dynamic graph convolution for semantic segmentation. In *European conference on computer vision* (pp. 1–17). Springer.

Huang, W., Chen, C., Li, Y., Li, J., Li, C., Song, F., et al. (2023). Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3061–3071).

Huang, J., Guan, D., Xiao, A., & Lu, S. (2021). Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6891–6902).

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 603–612).

Kim, J., Lee, J., Park, J., Min, D., & Sohn, K. (2022). Pin the memory: Learning to generalize semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4350–4360).

Lee, S., Seong, H., Lee, S., & Kim, E. (2022). WildNet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9936–9946).

Li, X., Li, M., Wang, Y., Ren, C.-X., & Guo, X. (2023). Adaptive texture filtering for single-domain generalized segmentation. In *Proceedings of the AAAI conference on artificial intelligence*.

Li, X., Li, X., You, A., Zhang, L., Cheng, G., Yang, K., et al. (2021). Towards efficient scene understanding via squeeze reasoning. *IEEE Transactions on Image Processing*, *30*, 7050–7063.

Liao, M., Hua, G., Tian, S., Zhang, Y., Zou, W., & Li, X. (2022). Exploring more concentrated and Consistent Activation Regions for cross-domain semantic segmentation. *Neurocomputing*.

Liao, M., Tian, S., Zhang, Y., Hua, G., Zou, W., & Li, X. (2023a). Calibration-based dual prototypical contrastive learning approach for domain generalization semantic segmentation. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 2199–2210).

Liao, M., Tian, S., Zhang, Y., Hua, G., Zou, W., & Li, X. (2023b). Domain-invariant information aggregation for domain generalization semantic segmentation. *Neurocomputing*, Article 126273.

Liao, M., Tian, S., Zhang, Y., Hua, G., Zou, W., & Li, X. (2024a). Calibration-based multi-prototype contrastive learning for domain generalization semantic segmentation in traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 1–17.

Liao, M., Tian, S., Zhang, Y., Hua, G., Zou, W., & Li, X. (2024b). PDA: Progressive domain adaptation for semantic segmentation. *Knowledge-Based Systems*, *284*, Article 111179.

Liao, M., Tian, S., Zhang, Y., Hua, G., Zou, W., & Li, X. (2024c). Preserving label-related domain-specific information for cross-domain semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 1–15.

Liu, B., Adeli, E., Cao, Z., Lee, K.-H., Shenoi, A., Gaidon, A., et al. (2020). Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, *5*(2), 3485–3492.

Liu, M., Schonfeld, D., & Tang, W. (2021). Exploit visual dependency relations for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9726–9735).

Ma, Y., Guo, Y., Liu, H., Lei, Y., & Wen, G. (2020). Global context reasoning for semantic segmentation of 3D point clouds. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2931–2940).

Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision* (pp. 116–131).

Neuhold, G., Ollmann, T., Rota Bulo, S., & Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision* (pp. 4990–4999).

Nguyen, B. X., Do, T., Tran, H., Tjiputra, E., Tran, Q. D., & Nguyen, A. (2022). Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4558–4566).

Niu, R., Sun, X., Tian, Y., Diao, W., Feng, Y., & Fu, K. (2021). Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–18.

Paithane, P., & Kakarwal, S. (2023). LMNS-net: Lightweight multiscale novel semantic-net deep learning approach used for automatic pancreas image segmentation in CT scan images. *Expert Systems with Applications*, *234*, Article 121064.

Pan, X., Luo, P., Shi, J., & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European conference on computer vision* (pp. 464–479).

Pan, X., Zhan, X., Shi, J., Tang, X., & Luo, P. (2019). Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1863–1871).

Peng, D., Lei, Y., Hayat, M., Guo, Y., & Li, W. (2022). Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2594–2605).

Peng, D., Lei, Y., Liu, L., Zhang, P., & Liu, J. (2021). Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, *30*, 6594–6608.

Qi, X., Liao, R., Jia, J., Fidler, S., & Urtasun, R. (2017). 3D graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 5199–5208).

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European conference on computer vision* (pp. 102–118). Springer.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). 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* (pp. 3234–3243).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., & Ré, C. (2020). No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, *33*, 19339–19352.

Suo, C., Zhou, T., Hu, K., Zhang, Y., & Gao, X. (2024). Cross-level collaborative context-aware framework for medical image segmentation. *Expert Systems with Applications*, *236*, Article 121319.

Tang, Q., Liu, F., Zhang, T., Jiang, J., Zhang, Y., Zhu, B., et al. (2022). Compensating for local ambiguity with encoder-decoder in urban scene segmentation. *IEEE Transactions on Intelligent Transportation Systems*.

Tjio, G., Liu, P., Zhou, J. T., & Goh, R. S. M. (2022). Adversarial semantic hallucination for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 318–327).

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6924–6932).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, J., Bao, B.-K., & Xu, C. (2021). Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, *24*, 3369–3380.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).

Wu, Y., Zhang, G., Gao, Y., Deng, X., Gong, K., Liang, X., et al. (2020). Bidirectional graph reasoning network for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9080–9089).

Xian, P., Po, L.-M., Xiong, J., Zhou, C., Zhao, Y., Yu, W.-Y., et al. (2022). Pixel voting decoder: A novel decoder that regresses pixel relationships for segmentation. *Expert Systems with Applications*, *193*, Article 116438.

Xie, G.-S., Liu, J., Xiong, H., & Shao, L. (2021). Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5475–5484).

Xu, Q., Yao, L., Jiang, Z., Jiang, G., Chu, W., Han, W., et al. (2022). DIRL: Domain-invariant representation learning for generalizable semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*.

Yang, L., Gu, X., & Sun, J. (2023). Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., et al. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2636–2645).

Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., & Gong, B. (2019). 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* (pp. 2100–2110).

Zeng, Z., Xu, Y., Xie, Z., Tang, W., Wan, J., & Wu, W. (2024). Large-scale point cloud semantic segmentation via local perception and global descriptor vector. *Expert Systems with Applications*, *246*, Article 123269.

Zhang, Y., Jiang, M., & Zhao, Q. (2021). Explicit knowledge incorporation for visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1356–1365).

Zhang, Y., Tian, S., Liao, M., Hua, G., Zou, W., & Xu, C. (2023). Learning shape-invariant representation for generalizable semantic segmentation. *IEEE Transactions on Image Processing*, http://dx.doi.org/10.1109/TIP.2023.3287506, 1–1.

Zhang, Y., Tian, S., Liao, M., Zhang, Z., Zou, W., & Xu, C. (2023). Fine-grained self-supervision for generalizable semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, http://dx.doi.org/10.1109/TCSVT.2023.3285091, 1–1.

Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., & Lee, G. H. (2022). Style-hallucinated dual consistency learning for domain generalized semantic segmentation. *European Conference on Computer Vision*.

Zhou, T., Li, L., Li, X., Feng, C.-M., Li, J., & Shao, L. (2021). Group-wise learning for weakly supervised semantic segmentation. *IEEE Transactions on Image Processing*, *31*, 799–811.

Muxin Liao: Currently, I am a lecturer in the School of Computer and Information Engineering at Jiangxi Agricultural University. Before that, I obtained my Ph.D. at the School of Electronics and Information Engineering, Shenzhen University in June 2024. My general research interest is to develop deep learning techniques for computer vision tasks. I mainly investigate how to exploit transferable representations and utilize them to develop domain adaptation and domain generalization semantic segmentation. Recently, I have focused on domain adaptation, domain generalization, semantic segmentation, and agricultural intelligence.