ProGBA: Prompt Guided Bayesian Augmentation for Zero-shot Domain Adaptation

Jian Zou[1], Guanglei Yang[1](✉), Tao Luo[2], Chun-Mei Feng[2], and Wangmeng Zuo[1]

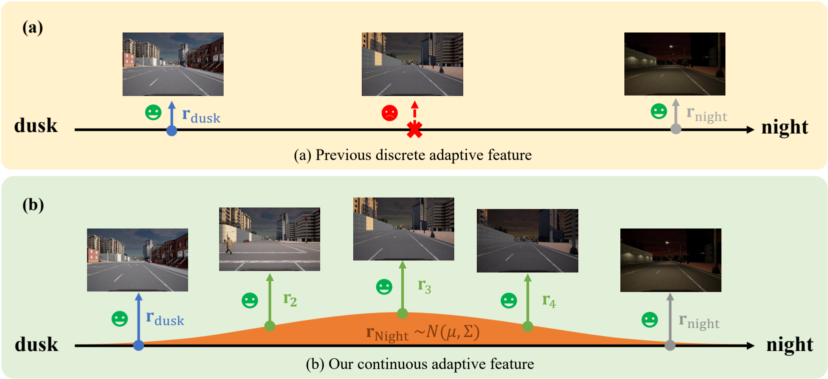
[1] Harbin Institute of Technology, China
yangguanglei@hit.edu.cn
[2] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore

Abstract. Domain adaptation is a well-established field within computer vision. Due to the common scenario of inaccessible target domain data, zero-shot domain adaptation increasingly gets more attention. Existing methods, which primarily focus on optimizing an Empirical Risk Minimization objective, tend to rely on training with discrete augmentations based on limited prompts. This strategy struggles to fully capture the complexity of the target domain, consequently diminishing the transferred model's effectiveness. In this paper, we introduce ProGBA, a novel framework that adopts a Bayesian perspective to regard the learning process in zero-shot domain adaptation as a variational inference problem. This approach aims to comprehend the distribution of domain-adaptive augmentations. Leveraging Bayesian methods' regularization capabilities, ProGBA refines the domain adaptation representation space, which helps to mitigate the overfitting risks. Specifically, ProGBA adeptly introduces the uncertainties associated with domain shifts through probabilistic modeling of residuals between the source and target domains, which reduces the model's reliance on a specific set of weights, thereby enhancing performance in the target domain. Furthermore, we adopt a pretrained visual-language model alongside a novel text-based loss function to more accurately align the learned distribution with the actual residual distribution between the target and source domains. The comprehensive validation showcases ProGBA's potential to set a new benchmark in zero-shot domain adaptation, demonstrating ProGBA's efficacy in adapting to the target domain. Moreover, extensive experiments on cross-domain semantic segmentation also underscore our method's generalizability.

Keywords: Zero-shot domain adaptation · bayesian learning · prompt guided

1 Introduction

In recent years, supervised methods for semantic segmentation [6, 11, 48, 80] and detection [3, 45, 78] have undergone significant advancements in both performance and computational efficiency [74, 87]. However, these improvements are


(a)
dusk
$\mathbf{r}_{\text{dusk}}$
$\mathbf{r}_{\text{night}}$
night
(b)
dusk
$\mathbf{r}_{\text{dusk}}$
$\mathbf{r}_2$
$\mathbf{r}_3$
$\mathbf{r}_4$
$\mathbf{r}_{\text{Night}} \sim N(\mu, \Sigma)$
$\mathbf{r}_{\text{night}}$
night

(a) Previous discrete adaptive feature

(b) Our continuous adaptive feature

Fig. 1: (a): Due to the learned discrete augmentations, methods based on the optimization of the Empirical Minimization objective fail when evaluated in the unpredefined target domain, even if it is similar to the existing target domain; **(b)**: ProGBA describes the augmentation in terms of distribution to describe the domain shifts and represent domain shifts within a range. For ease of presentation, we simplify the adaptation feature space to a single line.

contingent upon the stringent requirement that the context where the data used for both training and testing originates from the same domain. The efficacy of these methods sharply declines when applied to data from outside their training distribution [55], especially unseen domains. This significant performance degradation not only raises security concerns but also questions their reliability for critical real-world applications, such as autonomous driving, where safety is of paramount importance.

To alleviate the negative impact of domain drifts [2], a special case of unsupervised domain adaptation where target-domain samples are unavailable for training, named zero-shot domain adaptation, has gradually attracted attention. Given the challenge of zero-shot domain adaptation, the exploration of leveraging prior knowledge to synthetically generate data representative of the target domain becomes crucial [22,68,90]. At the beginning, [33,42,83] focus on establishing learnable parameters or modules to synthesize data for the target domain by integrating source data and description of other perspectives of target domain. Recently, Clip the Gap [72] utilizes a pre-trained vision-language model to infuse semantic concepts from the target domain through textual prompts, while Zhang *et al.* [42] exploit a visual inductive prior derived from physics-based reflection models for domain adaptation. Nonetheless, these methods primarily aim to optimize an Empirical Risk Minimization(ERM) objective and constrained to learning discrete adaptation features for a narrow range of predefined target domains, facing difficulties in adapting to the dynamic real-world environments. As depicted in Fig. 1 (a), simply sampling augmentation between *dusk* and *night* in the feature space fails to accurately represent a *twilight* feature for adaptation, often resulting in features with limited practical relevance.

To overcome the challenges previously discussed, this study introduces the ProGBA, a novel approach that adopts a Bayesian perspective for domain adaptation learning. ProGBA formulates the adaptation process as a variational inference problem, where the learning of domain adaptation is guided by principles of Bayesian inference. By leveraging textual descriptions, ProGBA integrates target domain information, modeling domain shifts from source to target as distributions — effectively treating these shifts as a form of data augmentation for the source domain. Intuitively, the random sampling from this distribution incorporates variability, thereby embedding uncertainty into the process. This uncertainty fosters a more robust representation of the latent space, as it compels the model to accommodate and depict the inherent variability in data. Consequently, this strategy prevents the fixation on overly specific adaptation parameters for unseen target domains, facilitating a broader, more generalized representation. Furthermore, ProGBA's approach to modeling augmentations through distributions replaces traditional discrete encodings with a continuous, more efficient representation, thus broadening the augmentation feature space coverage. This method, illustrated in Fig. 1 (b), enables ProGBA to learn an averaged representation of domain shifts. Moving towards this mean representation enhances the model's adaptability to the intended domain, a feat not attainable through discrete learning methods alone. Additionally, to ensure the augmentation distribution accurately reflects the true domain shifts from source to target, a novel loss function based on the Evidence Lower Bound (ELBO) is introduced. This loss function facilitates a closer approximation of the actual domain shifts, enhancing the model's effectiveness in domain adaptation.

To sum up, our contributions can be presented as follows:

- We frame zero-shot domain adaptation learning from the Bayesian perspective and model the domain shifts from source to target as a distribution, reducing the risk of overfitting and learning a more generalized augmentation across domains.
- A novel loss based on the ELBO is proposed to closely approximate the actual domain shifts from the source to the target domain.
- Comprehensive experiments demonstrate ProGBA's potential to promote the performance of zero-shot domain adaptation, showing its adaptability and generalizability on semantic segmentation.

2 Related Work

2.1 Domain adaptation

Due to its relevance in many practical applications, domain adaptation for semantic segmentation has been widely studied in the last few years. Hoffman *et al.* [26] first proposed a fully convolutional network for domain adaptive semantic segmentation integrating a Mutual Information loss and Generative Adversarial Networks (GANs) to realize domain feature alignment. Several works [8,9,34,52, 70,73] proposed to resort on adversarial learning and used a domain discriminator to alleviate the domain shift. Other approaches [10,20,25,32,66,67,75,76,79,91] adopted an image translation strategy and employed a GAN network to generate target-style images given an annotated source image. Another category of methods [4, 38, 41, 43, 49, 63, 77, 86] proposed to reduce the domain distribution mismatch by optimizing some divergence measures. More recent approaches [23, 29, 37, 44, 85, 88, 92] considered self-training to get supervision of the target texture. The above methods are effective under independent and identically distributed conditions, such as a synthetic-to-real or a cross-city. As a consequence, their performances typically degrade in the presence of significant domain gap [64]. To improve model robustness against domain gap, the latest methods exploit the potential of Transformers for unsupervised domain adaptation (UDA) semantic segmentation [27,28]. Furthermore, HGFormer [14] proposes a novel hierarchical grouping transformer, which makes Domain Adaptation Semantic Segmentation become classifications on all mask proposals. Nevertheless, in the industrial context [15, 51, 54], obtaining data for uncommon or rare target domains *e.g.* typhoons and sandstorm, presents a significant challenge. This issue is compounded by the high standards required for data quality and the challenges in ensuring the reliability of data sourced from the internet. Consequently, there's a need for domain adaptation techniques that can operate without actual target domain data, known as zero-shot domain adaptation.

2.2 Zero-shot domain adaptation

Zero-shot domain adaptation increasingly raise interest recently. Some works [33, 42] leverage extra information to accurately identify domain shifts that are applicable to the target domain. For instance, Yang *et al.* [83] employ data from multiple sources and target domains characterized by continuous variable vectors, while Ishii *et al.* [33] incorporate known attributes of the target domain *e.g.* the pose and position of a camera. ZDDA [59] deviates by using additional, unrelated source and target domain data pairs to better cater to domains relevant to the task of interest. This method aids in aligning the representations of source and target domains, yet its practical application is challenging. Moreover, domain generalization methods [12,30,58,72,89], which do not presuppose knowledge of the target domain, are deemed too broad to yield satisfying performance in specific target domains. Alternative strategies involve models that seek to directly learn the domain shift [35,50,72]. ULDA [82] enables efficient adaptation to multiple target domains. Their framework ensures semantic integrity by aligning features at scene, region, and pixel levels and maintaining relational consistency between visual representations and text embeddings. PØDA [18] marks a novel direction by employing a pre-trained CLIP [60] for zero-shot domain adaptation, using text descriptions of the target domain *i.e.* prompts, via the PIN module to adapt a source-trained model to target domains. Nevertheless, previous methods focus on optimizing an Empirical Risk Minimization objective, which tend to overfit to the source domain, thereby hampering performance. In contrast, our ProGBA leverages Bayesian methods based on prompts to regularize the feature space of augmentation. This not only prevents overfitting but

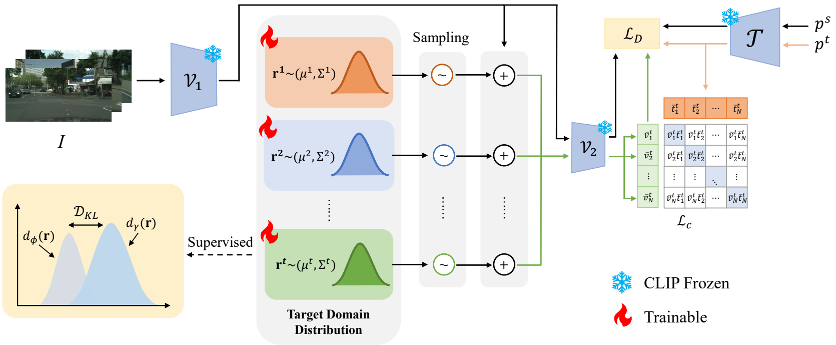

I
$\mathcal{V}_1$
$\mathbf{r}^1 \sim (\mu^1, \Sigma^1)$
$\mathbf{r}^2 \sim (\mu^2, \Sigma^2)$
$\mathbf{r}^t \sim (\mu^t, \Sigma^t)$
Target Domain Distribution
Sampling
$\mathcal{V}_2$
$\mathcal{L}_D$
$\mathcal{T}$
p^s
p^t
$\bar{t}_1^t$ $\bar{t}_2^t$ ⋯ $\bar{t}_N^t$
$\bar{v}_1^t$ $\bar{v}_2^t$ ⋮ $\bar{v}_N^t$
$\bar{v}_1^t\bar{t}_1^t$ $\bar{v}_1^t\bar{t}_2^t$ ⋯ $\bar{v}_1^t\bar{t}_N^t$
$\bar{v}_2^t\bar{t}_1^t$ $\bar{v}_2^t\bar{t}_2^t$ ⋯ $\bar{v}_2^t\bar{t}_N^t$
$\bar{v}_N^t\bar{t}_1^t$ $\bar{v}_N^t\bar{t}_2^t$ ⋯ $\bar{v}_N^t\bar{t}_N^t$
$\mathcal{L}_c$
$\mathcal{D}_{KL}$
$d_\phi(\mathbf{r})$
$d_\gamma(\mathbf{r})$
Supervised
CLIP Frozen
Trainable

Fig. 2: Overview of ProGBA. ProGBA initially estimates the distribution augmentation $\mathbf{r}$ for each the target domain, using textual domain prompts $(\mathcal{P}^t, p^s)$ and images from the source domain. This semantic enhancement aims to convert the embeddings of source domain images to match the target domain indicated by the prompts. During this distribution optimization phase, the loss $\mathcal{L}_O$ is minimized.

also provides a more nuanced understanding of the discrepancies between the source and target domains through a distribution-based approach.

2.3 Variational inference-based domain adaptation

Variational inference-based methods have proven effective in addressing domain adaptation and generalization challenges, leveraging the strengths of variational inference to enhance model robustness across different domains. MVI [5] aligns source and target domains in an indirect approach by maximizing mutual information between latent variables. Jing *et al.* [36] focus on source-free domain adaptation, using variational inference to generate perturbations that improve model robustness without accessing source data.Some other works [47] leverage variational inference to learn domain-invariant latent features. SIG [46] uses variational inference to identify a shared subspace, aligning multiple source domains with the target domain. MAVBI [19] combines variational inference with adversarial training to enhance generalization across multiple domains. However, these methods primarily concentrate on invariant features of source and target domain data or enhance model generalization [84] to adapt to the target domain, without explicitly modeling the target domain data. This limitation affects their performance in zero-shot domain adaptation. In contrast, ProGBA directly learns the residuals between the source and target domains to better adapt to the target domain, making it more suitable for zero-shot domain adaptation.

3 Proposed Method

3.1 Architecture

Distribution optimization In the context of zero-shot domain adaptation, our method is constrained to utilizing images exclusively from the source domain.

Our objective is to develop representations that remain effective despite variations in domain, achieved by simulating these domain shifts during the training phase on source data. To this end, ProGBA leverages the joint representation capability of CLIP [60] to approximate domain shifts within the visual domain by employing textual prompts. This process is depicted at Fig. 2.

In a formal setting, $\mathcal{T}$ is defined as the text encoder of CLIP and $\mathcal{V}$ as its image encoder. For clarity in our discussion, we further decompose $\mathcal{V}$ into two components: $\mathcal{V}_1$, which acts as a feature extractor, and $\mathcal{V}_2$, which serves as a projector into the CLIP embedding space, referring to [72]. The core training objective of CLIP is to align image features closely with their corresponding text prompts. This alignment is quantitatively measured by minimizing the distance between the embedded image features $\mathcal{V}_2(\mathcal{V}_1(I))$ and the text encoding $\mathcal{T}(p)$, for a given image I and its text prompt p. In the scenario of zero-shot domain adaptation, either directly or indirectly acquiring information about the target domain is essential. Fortunately, the linear word analogy property [17, 53] inherent in text embeddings allows for the use of algebraic operations to derive semantically similar concepts (*e.g.* the vector operation akin to *king* - *man* + *woman* closely aligns with the embedding for *queen*). This property of forming semantic relationships is also applicable to CLIP embeddings [61], enhancing its utility in domain adaptation tasks.

Leveraging the characteristic of word vectors, a relevant prompt p^s is predefined for the source domain, such as *Driving in the daytime*, and establish a corresponding prompt $\mathcal{P}^t = \{p^t_{i=1}\}^T_1$, where T stands for the number of target domains, for target domains *e.g. Driving at night.* ProGBA subsequently conceptualizes the augmentation space probabilistically as a distribution d_γ, anticipating that the bias introduced by $\mathbf{r}$ within d_γ mirrors the semantic divergence between the source prompt p^s and the target prompt $\mathcal{P}^t$. Orientedally, a learnable probability distribution d_γ acts on the features $F = \mathcal{V}_1(I)$ derived from the intermediate layer of the backbone. Inspired from the property of text embeddings mentioned above, we assume that the intermediate visual embedding within the target domain comprises two components: a fixed visual embedding F from the source domain, and a residual augmentation $\mathbf{r}$, which serves as a latent variable over the target visual embedding. Based on such hypothesis, ProGBA learns the latent distribution d_γ over the augmentation $\mathbf{r}$, *i.e.* $\mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are parameterized by two learnable vectors. Consequently, the enhanced features can be expressed with $\mathbf{r}$:

$$\bar{v}^t = \mathcal{V}_2\left(F + \mathbf{r}\right), \mathbf{r} \sim d_\gamma. \tag{1}$$

to approximate the *synthetic* visual embedding

$$\bar{v}' = v^s + \frac{\mathcal{T}\left(p^t\right) - \mathcal{T}\left(p^s\right)}{\left\|\mathcal{T}\left(p^t\right) - \mathcal{T}\left(p^s\right)\right\|_2}, v^s = \mathcal{V}_2\left(F\right) \tag{2}$$

in the target domain through the textual residual of the text embeddings. Ultimately, the augmentation distributions $D_\gamma = \{d^T_\gamma\}^T_{t=1}$ of different target domians are optimized through the loss function introduced in section 3.2.

Training with augmentation The adaptation strategy is aligned with [72]. When adapting to semantic segmentation, we adopt the following approach during each training iteration: the backbone network, initialized with pre-trained CLIP [60], is kept frozen, while the model's head is trained utilizing data from the source domain, which is enhanced by the optimized distribution of the target domain. Specifically, for each chosen target domain, ProGBA generates L Monte-Carlo samples from the target domain's augmentation distribution. These samples are then averaged to produce $\mathbf{r}$, which is utilized to train the model's head. Similar to the distribution optimization phase, the augmentation $\mathbf{r}$ is added to the intermediate layer outputs F of the backbone network. This output is then processed through the remaining layers of the backbone and the task-oriented head, encapsulating the process as follows:

$$\bar{F} = \mathcal{V}_2\left(\mathcal{V}_1(I) + \frac{1}{L}\sum_{1}^{L}\mathbf{r}_l\right), \mathbf{r}_l \sim d_\gamma. \tag{3}$$

By enhancing the backbone network with low-level feature augmentations, the high-level semantic content or features of the image avoid distortion. Consequently, the original labels from the source domain are still applicable for training purposes without necessitating any alterations. This characteristic facilitates the application of ProGBA to various models with minimal adjustments required. In the case of segmentation tasks, ProGBA is adapted for use with SAN [80], a segmentation model that incorporates CLIP for classification purposes. Throughout the training process, a specifically designed loss function is incorporated, which will be elaborated upon subsequently, to ensure comprehensive coverage of the target domain by the augmentation distribution. At the inference stage, the model is deployed without applying any augmentation distribution, thereby not introducing any additional complexity nor affecting the model's inference speed.

3.2 Loss function

To achieve an augmentation $\mathbf{r}$ that closely approximates the actual distribution of residuals between the target and source domains, we define our optimization goal for a set of N annotated data $\{\mathbf{x}_i = (I_i, \mathcal{P}^t, p^s), \mathbf{y}_i\}_{i=1}^N$ as follows:

$$\mathbf{r}^* = \arg\min_{\mathbf{r}} \mathbb{E}_{\mathbf{x}_i,\mathbf{y}_i}\left[-\log P\left(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{r}\right)\right] \tag{4}$$

Denoting the augmented domain-specific feature as $\mathcal{V}_\gamma(I) = \mathcal{V}_2(\mathcal{V}_1(I) + \mathbf{r})$, the marginal likelihood P, referring to [1], is then defined as

$$P(\mathbf{y} \mid \mathbf{x}) = \int_\gamma \frac{e^{\mathcal{V}_\gamma(I)^T \mathcal{T}\left(p^t\right)}}{\sum_{t'} e^{\mathcal{V}_\gamma(I)^T \mathcal{T}\left(p^{t'}\right)}} P\left(\mathcal{V}_\gamma(I)\right) d\gamma \tag{5}$$

Addressing Eq. 4 using the marginal likelihood outlined in Eq. 5 proves to be unfeasible, requiring the indeterminable d_γ. Drawing inspiration from Variational Autoencoders (VAE) [40], our strategy pivots to establishing a lower bound through the introduction of a variational distribution d_ϕ, enabling the sampling of the residual $\mathbf{r}$. This variational bound is articulated as:

$$\log P(\mathbf{y} \mid \mathbf{x}) \geq \mathbb{E}_{d_\phi(\mathbf{r})}[\log P(\mathbf{y} \mid \mathbf{x}, \mathbf{r})] - \mathcal{D}_{\mathrm{KL}}\left[d_\phi(\mathbf{r}) \| d_\gamma(\mathbf{r})\right] \tag{6}$$

with $\log P(\mathbf{y} \mid \mathbf{x}, \mathbf{r}) \propto e^{\mathcal{V}_\gamma(I)^T \mathcal{T}(p^t)}$, where the dependency on $\mathbf{r}$ originates from the definition of $\mathcal{V}_\gamma(I)$. To maximize the expectation $\mathbb{E}_{d_\phi(\mathbf{r})}[\log P(\mathbf{y} \mid \mathbf{x}, \mathbf{r})]$ in Eq. 6, ProGBA minimizes the loss:

$$\mathcal{L}_c = -\frac{1}{2N} \sum_{i=1}^{N} \left(\log \frac{\exp\left(t_i v_i / \tau\right)}{\sum_{j=1}^{N} \exp\left(t_i v_j / \tau\right)} + \log \frac{\exp\left(t_i v_i / \tau\right)}{\sum_{j=1}^{N} \exp\left(t_j v_i / \tau\right)} \right) \tag{7}$$

where τ is a temperature parameter referring to [60].

Adhering to [21, 40], we designate d_ϕ as a learnable Gaussian distribution, as $\mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ being vectors subject to learning. $d_\gamma(\mathbf{r})$ is defined as a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathrm{I})$. Subsequently, a reparameterization method is used to generate Monte Carlo samples from d_ϕ to minimize $\mathcal{D}_{\mathrm{KL}}$ in Eq. 6. Consequently, our proposed loss function can be articulated as:

$$\mathcal{L}_{bayes} = \mathcal{L}_c + \mathcal{D}_{\mathrm{KL}}\left[d_\phi(\mathbf{r}) \| d_\gamma(\mathbf{r})\right] \tag{8}$$

Inspired from [72], the loss function during the distribution optimization stage further involves the cosine distance between $\bar{v}'$ and $\bar{v}^t$, which is formulated as:

$$\mathcal{L}_D = \mathcal{D}\left(\bar{v}', \bar{v}^t\right) + \|v^s - \bar{v}^t\|_1, \mathcal{D}(x, y) = 1 - \frac{x - y}{\|x - y\|_2} \tag{9}$$

An l_1 regularizer is utilized to ensure embeddings do not stray significantly from their original values, thereby maintaining the integrity of image semantic content. The overall loss $\mathcal{L}_O$ in the distribution optimization stage can be expressed as:

$$\mathcal{L}_O = \mathcal{L}_{bayes} + \mathcal{L}_D \tag{10}$$

As for the loss during training with augmentation, $\mathcal{D}_{\mathrm{KL}}$ is also involved:

$$\mathcal{L}_T = \lambda \mathcal{D}_{\mathrm{KL}}\left[d_\phi(\mathbf{r}) \| d_\gamma(\mathbf{r})\right] + \mathcal{L}_{spec} \tag{11}$$

where $\mathcal{L}_{spec}$ denotes the task-oriented loss. λ is set to 0.01 by default, as determined by the ablation experiments in the Suppl.

4 Experiments

4.1 Experimental setup

Datasets We utilize the Cityscapes dataset [13] as our primary source in semantic segmentation, which consists of 2,975 training images and 500 validation images, spanning 19 semantic classes. However, since the validation set of

Cityscapes is divided by cities like the training set, which is not applicable to test domain adaptation methods due to its insignificant domain gap with the training set, our primary results are benchmarked using the ACDC dataset [65], selected for its collection of urban images under challenging weather conditions. Furthermore, we explore the generalizability of ProGBA across two distinct scenarios: transferring from real to synthetic environments (using Cityscapes as the source and GTA5 [62] for evaluation) and from synthetic to real (with GTA5 as the source and Cityscapes for assessment). Evaluations are conducted on the provided validation sets. For GTA5, a randomly selected subset of 1,000 images is used for analysis, referring to the settings of PØDA [18].

Adapting to continuously evolving environments represents a safety-critical challenge for zero-shot domain adaptation methods. To assess the efficacy of ProGBA in dynamic scenarios, we employed the SHIFT dataset [69]. SHIFT is a synthetic dataset designed for autonomous driving, characterized by its unique representation of continuous environmental variations, including changes in cloudiness, precipitation, fog intensity, and time of day. The dataset comprises 4,850 sequences, with each sequence consisting of 500 frames recorded at a frequency of 10 Hz. For a more targeted evaluation, we reorganized the validation set into three subsets—Night, Fog, and Rain—based on the type of domain shift, and subsequently assessed the performance of ProGBA on these subsets.

Evaluation metrics Our experiments utilize the Mean of class-wise Intersection over union (mIoU) metric to measure the performance of adaptation in semantic segmentation. For PØDA and ProGBA, the mean and standard deviation are reported over three models trained with different random seeds.

4.2 Implementation details

Distribution optimization As the benchmark dataset evaluates semantic segmentation performance across various weather conditions, a collection of target domain prompts $\mathcal{P}^t$ are crafted to align with the validation split's weather conditions, comprising snow, fog, rain, and sunshine, alongside three distinct times of the day: day, night, and evening. The combination of sunshine are excluded during the daytime from our prompts since it corresponds to the source domain. This setup enables us to create $M = 11$ distinct prompts following the format *an image taken on a* $\{$*weather*$\}$ $\{$*time of the day*$\}$. For the source domain prompt p^s, we use *an image taken during the day.* The prompts used in Real→Synthetic, Synthetic→Real in semantic segmentation are designed similarly.

To finetune the augmentations with these prompts, we generate random crops from the source images, resizing them to 224 × 224 pixels. The mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ are set to 0 and 1 initially. The image encoder $\mathcal{V}$ is implemented as a ViT-B/16 [16], and the text encoder $\mathcal{T}$ employs the transformer architecture [71], both initialized with weights pre-trained using CLIP [60] and maintained frozen throughout the training process. For the optimization phase, we use the Adam optimizer [39], conducting the optimization over 2500 iterations with a learning rate of 1e-4.

Table 1: Zero-shot domain adaptation results on semantic segmentation. The mIoU scores grouped by source domain and target domain are reported. Source-only: SAN [80] which solely initialized using CLIP pre-training weights, without any further domain adaptation. *: Our re-implementation with visual backbone ViT-B/16.

Source	Target eval	Method	mIoU
Cityscapes	Target domain = Night		
	ACDC Night	Source-only [80]	26.36
		PØDA* [18]	28.24 ± 0.47
		ProGBA	**30.18** ± 0.39
	Target domain = Snow		
	ACDC Snow	Source-only [80]	47.47
		PØDA* [18]	48.61 ± 0.41
		ProGBA	**48.94** ± 0.34
	Target domain = Rain		
	ACDC Rain	Source-only [80]	47.64
		PØDA* [18]	47.89 ± 0.45
		ProGBA	**49.65** ± 0.37
	Target domain = Fog		
	ACDC Fog	Source-only [80]	59.83
		PØDA* [18]	59.84 ± 0.49
		ProGBA	**60.34** ± 0.35
	Target domain = Synthesis		
	GTA5	Source-only [80]	46.27
		PØDA* [18]	46.46 ± 0.43
		ProGBA	**47.69** ± 0.36
GTA5	Target domain = Real		
	Cityscapes	Source-only [80]	43.13
		PØDA* [18]	43.55 ± 0.44
		ProGBA	**43.67** ± 0.30

Training with augmentation While training with augmentation, we largely adhere to the SAN [80] with only minor adjustments to accommodate the adaptive segmenter. We process the input images by randomly resizing them to fall within the short-side range of $[320, 1024]$ before cropping them to $[640, 640]$. Both the visual encoder $\mathcal{V}$ and text $\mathcal{T}$ encoders are initialized using weights from the pre-training weights of CLIP [60] of ViT-B [16] version. Throughout the training phase, we maintain the text encoder $\mathcal{T}$ and visual encoder $\mathcal{V}$ in a frozen state, focusing on training the side adaptive network and the augmentation distribution. This training is conducted using the AdamW optimizer over 60,000 iterations, starting with an initial learning rate of 1e-4, a weight decay of 1e-4 and the batch size of 32. The learning rate follows a poly schedule with a decay power of 0.9.

4.3 Main results

ProGBA for semantic segmentation We compare performance of ProGBA, which utilizes semantic augmentation, with that of the state-of-the-art PØDA [18]. Additionally, we examine the efficiency of source-only as the baseline. Focusing on the semantic segmentation task, we train all model on Cityscapes and assesse the adaptability in Table 1 to unseen target domains by evaluating mIoU scores across various out-of-domain datasets, including conditions of night, rain, snow, and fog. Our ProGBA, which integrates CLIP pre-training model

Table 2: Zero-shot domain adaptation results on the continuous scenarios. Performance grouped by target domain are reported in terms of mIoU scores. Source-only: SAN [80] which solely initialized using CLIP pre-training weights, without any further domain adaptation. *: Our re-implementation with visual backbone ViT-B/16.

Target eval	Method	mIoU
	Target domain = Night	
SHIFT Night	Source-only [80]	39.82
	PØDA* [18]	41.88 ± 0.47
	ProGBA	**42.53** ± 0.35
	Target domain = Snow	
SHIFT Fog	Source-only [80]	43.19
	PØDA* [18]	44.54 ± 0.44
	ProGBA	**46.65** ± 0.42
	Target domain = Rain	
SHIFT Rain	Source-only [80]	36.89
	PØDA* [18]	37.58 ± 0.40
	ProGBA	**39.89** ± 0.31

with Bayesian prompt guided augmentation, demonstrates superior performance across all evaluated domains when compared to the source-only method. Despite not tailoring the learning process to individual images, our method achieves an improvement of approximate 2% mIoU over PØDA in challenging conditions like night and rain, while maintaining similar performance in snow and fog situations. This indicates the effective coverage of the learned distributions over the residual space between the source and target domains.

Transitioning from synthetic to real applications represents a common challenge in domain adaptation, especially in the autonomous driving. Since real labelled data in driving scenarios is often costly to obtain and data in an unusual weather *e.g. daytime foggy* is difficult to obtain, models are often trained on large amounts of synthetic data and eventually migrated to real scenarios. Therefore, minimizing the domain gap between synthetic and real environments is crucial. As presented in Table 1, ProGBA surpasses two baseline models by a margin of almost 1.2 mIoU, showcasing its effectiveness in bridging the domain gap. Moreover, when adapting Real → Synthetic, our method continues to outperform others, achieving the highest mIoU. Fig. 3 provides a qualitative results on different target domains. In conditions like CS → ACDC Rain, objects detected by PØDA are frequently partial, while ProGBA achieves more comprehensive segmentation for objects. As for CS → Synthesis, ProGBA more effectively adjusts to the target domain and accurately classifies objects.

ProGBA for continuous scenarios To further assess the generalizability of our approach, we conduct a comparative analysis with PØDA using the SHIFT synthetic dataset [69]. As shown in Table 2, the mean Intersection over Union (mIoU) is validated across three distinct domain shift types within the SHIFT dataset, utilizing Cityscapes as the training source domain. ProGBA demonstrates a mean improvement of 3.05 mIoU compared to the Source-only approach

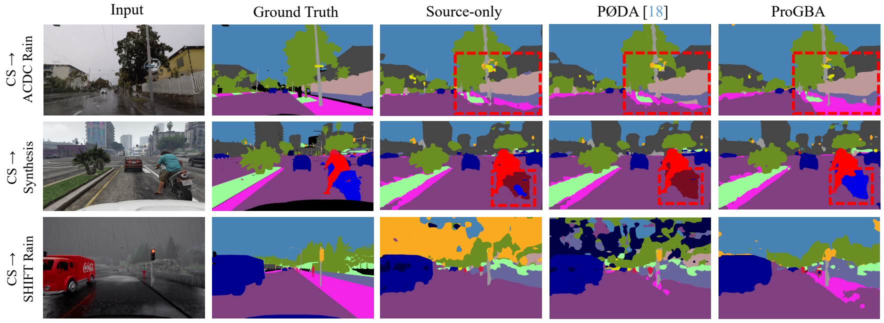


Fig. 3: Qualitative examples on Cityscapes → GTA5(Synthesis) validation set and Cityscapes → ACDC validation set.

across these scenarios, and an enhancement of up to 2.3 mIoU over PØDA, which similarly employs a clip-based methodology. These results underscore the effectiveness of the prompt-guided distribution learned by ProGBA in accommodating various levels of domain shifts. In contrast, PØDA utilizes a single optimal augmentation feature per sample, derived from a singular textual description, which may lead to overfitting at specific domain shift levels. Consequently, in scenarios characterized by continuous variations encompassing multiple domain shifts, PØDA achieves only an average mIoU improvement of 1.36%. The last row of Fig 3 presents the visualization results of Cityscapes → SHIFT validation set. Even in adverse weather conditions, such as heavy rain accompanied by fog, ProGBA effectively segments the sky, *etc*., demonstrating the robustness of our model in inclement weather.

4.4 Ablation study

Loss ablation The primary distinction between ProGBA and ERM-based domain adaptation approaches lies in the learning of distributions and the associated loss functions. Our loss ablation, detailed in Table 3, examines the effects of the component in our specifically designed loss $\mathcal{L}_{bayes}$. The data reveals that relying solely on $\mathcal{D}_{\mathrm{KL}}$ for guidance actually decreases mIoU scores in target domains characterized by fog and rain when compared to the methods where $\mathcal{D}_{\mathrm{KL}}$ is omitted. This decline in performance can be attributed to the fact that exclusively using $\mathcal{D}_{\mathrm{KL}}$, each augmentation tend to learn as a standard normal distribution, which inadequately represents the characteristics of the target domain. However, incorporating both $\mathcal{D}_{\mathrm{KL}}$ and $\mathcal{L}_c$ leads to a consistent mIoU improvement of approximately 1.6% across all target domains. This outcome validates the effectiveness of our proposed $\mathcal{L}_{bayes}$.

Selection of features to augment In Table 4, our analysis focuses on the impact of different feature layer selections for augmentation. The results evaluated on the ACDC night split indicate that optimal performance, marked by

Table 3: Ablation study of loss function. We study the influence for two different components of the loss function: Both distribution optimization and training with augmentation are performed on the Cityscapes train split, while the evaluation is conducted on the ACDC validation set.

		Cityscapes →				
$\mathcal{D}_{\mathrm{KL}}$	$\mathcal{L}_c$	ACDC Night	ACDC Fog	ACDC Rain	ACDC Snow	mIoU
		29.67	60.24	49.28	47.55	47.53
✓		29.55	59.57	**49.88**	47.69	48.45
✓	✓	**30.18**	**60.34**	49.64	**48.94**	**49.18**

Table 4: Impact of selection of features to augment. Performance on the ACDC Night validation set is reported.

Description	Layers	mIoU
single-adaption	3rd layer	28.16
	6th layer	**30.18**
	9th layer	27.71
multi-adaption	{3,6}-layer	27.98
	{3,9}-layer	26.29
	{6,9}-layer	27.08
	{3,6,9}-layer	27.00

Table 5: Performance with ResNet50 for DeepLabV3+ [7] head is reported. Models are validated on the ACDC Night split.

Method	mIoU
Source-only [7]	20.1
IBN-Net [56]	21.3
InterNorm [31]	23.8
SW [57]	20.5
RobustNet(ISW) [12]	23.2
SAN-SAW [58]	24.5
DPCL [81]	24.9
PØDA [18]	25.0
ULDA [82]	25.4
ProGBA	**25.5**

a 30.18 mIoU score, is achieved when augmenting features from the 6th layer of backbone. Moreover, it's observed that augmenting features across multiple layers yields inferior results compared to augmenting features from a single layer. This performance drop could stem from the increased complexity in optimizing the model when multiple layers are enhanced. Specifically, the increased optimization challenge when augmentation $\mathbf{r}$ from the same distribution is added to multiple backbone layers, *e.g.* 3^{rd} and 6^{th}, rather than single layer, *e.g.* 6^{th}. Moreover, $\mathbf{r}$ that sampled from the same distribution tends to be similar, which would introduce much redundancy and disrupt to original semantic information.

Other architecture In Table 5, we report the performance using ResNet50 [24] as a backbone in CS → ACDC Night. Compared to the Source-only, ProGBA achieves a significant improvement of 5.4 mIoU, outperforming other SOTA zero-shot domain adaptation methods and domain generalization approaches which are similar to the zero-shot domain adaptation setups. Notably, ProGBA demonstrates the consistent improvement with a very different backbone (ResNet50) compared to ViT [16], further underscoring the generalizability of our method.

Number of Monte Carlo sampling The quantity of Monte Carlo samples plays a crucial role in accurate log-likelihood approximation, with more samples generally leading to more accurate approximations and better adaptation to target domains. Table 6 delves into this aspect by altering the count of Monte Carlo samples within the learned normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We observed a consistent improvement as the number of samples is increased from 1 to 10,

Table 6: Impact of the Monte Carlo sampling on target domain. Mean: the average mIoU of CS→ACDC Night and CS→GTA5.

#Monte Carlo sampling	CS → ACDC Night	CS → GTA5	Mean
1	29.67	46.79	38.23
5	**30.20**	46.95	38.58
10	30.18	47.69	**38.94**
20	29.22	**47.86**	38.54

Table 7: Impact of the variational distribution on target domain. The average mIoU of CS→ACDC Night and CS→GTA5 is reported as Mean.

Variational distribution	CS → ACDC Night	CS → GTA5	Mean
$\mathcal{U}(0,1)$	29.14	46.78	36.96
$\mathcal{N}(\mathbf{0},\mathrm{I})$	29.66	47.15	38.41
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	**30.18**	**47.69**	**38.94**

beyond which the mIoU levels off. Consequently, we settled on 10 as the optimal number of Monte Carlo samples for our augmentation module.

Variational distribution We ablate the effectiveness of the variational distribution by comparing the impact of sampling augmentations from different distributions: the uniform distribution $\mathcal{U}(0,1)$, the normal distribution $\mathcal{N}(\mathbf{0},\mathrm{I})$, and the specifically tailored normal distribution $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. The results, detailed in Table 7, reveal that samples drawn from $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ consistently outperform those obtained from the uniform distribution $\mathcal{U}(0,1)$ and the generic normal distribution $\mathcal{N}(\mathbf{0},\mathrm{I})$. This superiority demonstrates the effectiveness of the enhancements produced by our prompt guided bayesian augmentation module, highlighting its ability to generate informative adaptive augmentations.

5 Conclusion

ProGBA introduces a novel approach for zero-shot domain adaptation by employing prompt-guided bayesian augmentation, framing the adaptation learning as a variational inference problem. Two principal innovations define this method: firstly, the domain shift from source to target is modeled as a distribution, reducing the risk of overfitting and learning more generalized augmentation across domains; and secondly, a novel loss based on ELBO is proposed to closely approximate the actual domain shifts. Extensive experiments reveal marked enhancements, highlighting ProGBA's adaptability and broad applicability.

Limitations The distribution parameters ProGBA uses are unconditional *i.e.* they are learned across the whole dataset and fixed during testing, which restricts the model's potential. Additionally, the dependency on the latent space of visual-language models means ProGBA cannot be applied to the backbone without large-scale pre-training, narrowing its applicability.

Acknowledgments This work was partially supported by National Key RD Program of China under Grant No. 2021ZD0112100. Guanglei Yang is supported by the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20233458. Tao Luo is supported by the A*STAR Funding under Grant Number C230917003.

References

1. Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R.A., Murphy, K.: Fixing a broken elbo. In: International conference on machine learning. pp. 159–168. PMLR (2018)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**, 151–175 (2010)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chang, W.L., Wang, H.P., Peng, W.H., Chiu, W.C.: All about structure: Adapting structural information across domains for boosting semantic segmentation. In: CVPR (2019)
5. Chen, J., Wang, J., de Silva, C.W.: Mutual variational inference: An indirect variational inference approach for unsupervised domain adaptation. IEEE Transactions on Cybernetics **52**(11), 11491–11503 (2021)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: ICCV (2017)
9. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: CVPR (2018)
10. Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B.: Crdoco: Pixel-level domain transfer with cross-domain consistency. In: CVPR (2019)
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
12. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11580–11590 (2021)
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
14. Ding, J., Xue, N., Xia, G.S., Schiele, B., Dai, D.: 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 (2023)
15. Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X.: Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. arXiv preprint arXiv:2401.00988 (2024)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is

worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

17. Ethayarajh, K., Duvenaud, D., Hirst, G.: Towards understanding linear word analogies. arXiv preprint arXiv:1810.04882 (2018)
18. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., De Charette, R.: Poda: Prompt-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18623–18633 (2023)
19. Gao, Z., Guo, S., Xu, C., Zhang, J., Gong, M., Del Ser, J., Li, S.: Multi-domain adversarial variational bayesian inference for domain generalization. IEEE Transactions on Circuits and Systems for Video Technology (2022)
20. Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. In: CVPR (2019)
21. Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., Turner, R.E.: Meta-learning probabilistic inference for prediction. arXiv preprint arXiv:1805.09921 (2018)
22. Gu, Q., Zhou, Q., Xu, M., Feng, Z., Cheng, G., Lu, X., Shi, J., Ma, L.: Pit: Position-invariant transform for cross-fov domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8761–8770 (2021)
23. He, J., Jia, X., Chen, S., Liu, J.: Multi-source domain adaptation with collaborative learning for semantic segmentation. In: CVPR (2021)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML. PMLR (2018)
26. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv (2016)
27. Hoyer, L., Dai, D., Van Gool, L.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: CVPR. pp. 9924–9935 (2022)
28. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: ECCV. pp. 372–391. Springer (2022)
29. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: CVPR (2023)
30. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6891–6902 (2021)
31. Huang, L., Zhou, Y., Zhu, F., Liu, L., Shao, L.: Iterative normalization: Beyond standardization towards efficient whitening. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4874–4883 (2019)
32. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. arXiv (2018)
33. Ishii, M., Takenouchi, T., Sugiyama, M.: Zero-shot domain adaptation based on attribute information. In: Lee, W.S., Suzuki, T. (eds.) Proceedings of The Eleventh Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 101, pp. 473–488. PMLR (17–19 Nov 2019), https://proceedings.mlr.press/v101/ishii19a.html
34. Isobe, T., Jia, X., Chen, S., He, J., Shi, Y., Liu, J., Lu, H., Wang, S.: Multi-target domain adaptation with collaborative consistency learning. In: CVPR (2021)

35. Jhoo, W.Y., Heo, J.P.: Collaborative learning with disentangled features for zero-shot domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8896–8905 (2021)
36. Jing, M., Zhen, X., Li, J., Snoek, C.: Variational model perturbation for source-free domain adaptation. Advances in Neural Information Processing Systems **35**, 17173–17187 (2022)
37. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: CVPR (2020)
38. Kim, Y., Lee, J., Park, C., won Kim, H., Lim, I., Chang, C., Choi, J.W.: Semi-supervised domain adaptation using target-oriented domain augmentation for 3d object detection. IEEE Transactions on Intelligent Vehicles (2024)
39. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
40. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
41. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced wasserstein discrepancy for unsupervised domain adaptation. In: CVPR (2019)
42. Lengyel, A., Garg, S., Milford, M., van Gemert, J.C.: Zero-shot day-night domain adaptation with a physics prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4399–4409 (2021)
43. Li, G., Kang, G., Liu, W., Wei, Y., Yang, Y.: Content-consistent matching for domain adaptive semantic segmentation. In: ECCV (2020)
44. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: CVPR (2019)
45. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022)
46. Li, Z., Cai, R., Chen, G., Sun, B., Hao, Z., Zhang, K.: Subspace identification for multi-source domain adaptation. Advances in Neural Information Processing Systems **36** (2024)
47. Liu, X., Hu, B., Jin, L., Han, X., Lu, F.X.J.O.J., Woo, G.E.F.J.: Rethinking the invariant feature learning: Variational bayesian inference for domain generalization. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. pp. 881–882 (2021)
48. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
49. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML (2015)
50. Luo, R., Wang, W., Yang, W., Liu, J.: Similarity min-max: Zero-shot day-night domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8104–8114 (2023)
51. Ma, Z., Yang, Y., Wang, G., Xu, X., Shen, H.T., Zhang, M.: Rethinking open-world object detection in autonomous driving scenarios. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1279–1288 (2022)
52. Michieli, U., Biasetton, M., Agresti, G., Zanuttigh, P.: Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation. IEEE Transactions on Intelligent Vehicles **5**(3), 508–518 (2020)
53. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

54. Mușat, V., Fursa, I., Newman, P., Cuzzolin, F., Bradley, A.: Multi-weather city: Adverse weather stacking for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2915 (2021)
55. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32** (2019)
56. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 464–479 (2018)
57. Pan, X., Zhan, X., Shi, J., Tang, X., Luo, P.: Switchable whitening for deep representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1863–1871 (2019)
58. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-aware domain generalized segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2594–2605 (2022)
59. Peng, K.C., Wu, Z., Ernst, J.: Zero-shot deep domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 764–781 (2018)
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
61. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125 **7** (2022)
62. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 102–118. Springer (2016)
63. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR (2018)
64. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV (2019)
65. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10765–10775 (2021)
66. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: CVPR (2018)
67. Schutera, M., Hussein, M., Abhau, J., Mikut, R., Reischl, M.: Night-to-day: Online image-to-image translation for object detection within autonomous driving by night. IEEE Transactions on Intelligent Vehicles **6**(3), 480–489 (2020)
68. Sindagi, V.A., Oza, P., Yasarla, R., Patel, V.M.: Prior-based domain adaptive object detection for hazy and rainy conditions. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 763–780. Springer (2020)
69. Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., Yu, F.: Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21371–21382 (2022)

70. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
71. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
72. Vidit, V., Engilberge, M., Salzmann, M.: Clip the gap: A single domain generalization approach for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3219–3229 (2023)
73. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019)
74. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: Yolov9: Learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616 (2024)
75. Wang, J., Jiang, J.: Conditional coupled generative adversarial networks for zero-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3375–3384 (2019)
76. Wang, J., Jiang, J.: Adversarial learning for zero-shot domain adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 329–344. Springer (2020)
77. Wen, L., Xu, Y., Feng, Z., Zhou, J., Zhou, L., Wang, Y.: Semi-supervised domain adaptation for semantic segmentation via active learning with feature-and semantic-level alignments. IEEE Transactions on Intelligent Vehicles (2024)
78. Wu, D., Liao, M.W., Zhang, W.T., Wang, X.G., Bai, X., Cheng, W.Q., Liu, W.Y.: Yolop: You only look once for panoptic driving perception. Machine Intelligence Research **19**(6), 550–562 (2022)
79. Wu, Z., Wang, X., Gonzalez, J.E., Goldstein, T., Davis, L.S.: Ace: Adapting to changing environments for semantic segmentation. In: ICCV (2019)
80. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
81. Yang, L., Gu, X., Sun, J.: Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 10789–10797 (2023)
82. Yang, S., Tian, Z., Jiang, L., Jia, J.: Unified language-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23407–23415 (2024)
83. Yang, Y., Hospedales, T.: Zero-shot domain adaptation via kernel regression on the grassmannian. arXiv preprint arXiv:1507.07830 (2015)
84. Zhang, L., Du, Y., Shen, J., Zhen, X.: Learning to learn with variational inference for cross-domain image classification. IEEE Transactions on Multimedia **25**, 3319–3328 (2022)
85. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR (2021)
86. Zhang, S., Zhang, L., Li, G., Li, P., Liu, Z.: Multi-prototype guided source-free domain adaptive object detection for autonomous driving. IEEE Transactions on Intelligent Vehicles (2023)
87. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European conference on computer vision (ECCV). pp. 405–420 (2018)

88. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. IJCV (2021)
89. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)
90. Zhou, Q., Gu, Q., Pang, J., Lu, X., Ma, L.: Self-adversarial disentangling for specific domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8954–8968 (2023)
91. Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D.: Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: ECCV (2018)
92. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018)