Towards Domain-Aware Knowledge Distillation for Continual Model Generalization

Nikhil Reddy[1] Mahsa Baktashmotlagh[2] Chetan Arora[3]
[1]UQ-IIT Delhi Research Academy (UQIDAR) [2]UQ, Australia [3]IIT Delhi, India
https://dose-iitd.github.io/

Abstract

*Generalization on unseen domains is critical for Deep Neural Networks (`DNNs`) to perform well in real-world applications such as autonomous navigation. However, catastrophic forgetting limits the ability of domain generalization and unsupervised domain adaption approaches to adapt to constantly changing target domains. To overcome these challenges, We propose `DoSe` framework, a **Do**main-aware **Se**lf-Distillation method based on batch normalization prototypes to facilitate continual model generalization across varying target domains. Specifically, we enforce the consistency of batch normalization statistics between two batches of images sampled from the same target domain distribution between the student and teacher models. To alleviate catastrophic forgetting, we introduce a novel exemplar-based replay buffer to identify difficult samples for the model to retain the knowledge. Specifically, we demonstrate that identifying difficult samples and updating the model periodically using them can help in preserving knowledge learned from previously seen domains. We conduct extensive experiments on two real-world datasets ACDC, C-Driving, and one synthetic dataset SHIFT to verify the efficiency of the proposed `DoSe` framework. On ACDC, our method outperforms existing `SOTA` in Domain Generalization, Unsupervised Domain Adaptation, and Daytime settings by 26%, 14%, and 70% respectively.*

1. Introduction

Generalization performance in unseen scenarios plays an important role in the performance of Deep Neural Networks (`DNNs`) in real-world applications such as autonomous navigation. `DNNs` have been shown to be robust against small shifts in the data distribution. However, when there is a substantial and prolonged alteration in the data distribution, referred to as the "Continuous shift", the model's performance on new scenarios may significantly decline. For example, an autonomous navigation system designed for clear weather conditions can perform significantly poor in low-light conditions.

Despite their success in urban-scene semantic segmentation task under ideal daytime conditions, the performance of `SOTA` techniques such as DeepLabv3+ ResNet101 [7] and DeepLabv2 MobileNet V2 [6], deteriorate significantly under challenging scenarios such as nighttime or adverse weather conditions. On the other hand, obtaining labeled ground truth data for training `DNN` models at night and in adverse weather conditions is difficult and costly.

Existing methods in Domain Generalization (`DG`), and Unsupervised Domain Adaptation (`UDA`) for semantic segmentation fail to perform well in a continually changing target environment due to phenomena referred to as catastrophic forgetting. To overcome this, Continual Domain Adaptation methods (`CDA`) [29, 50] aim to continually adapt a model to the unseen domains without degrading performance on previously seen domains. However, existing `CDA` methods require access to the source dataset, which limits its applicability due to limited storage and privacy concerns. This work aims to complement the aforementioned methods to adapt to the target domain with continually changing target distribution.

In a real-world setting, collecting (even unlabeled) target domain images for all adverse visual circumstances or their exponential number of combinations is not feasible. Moreover, due to the dynamic nature of the real world, the target data itself may change during inference. This motivates the development of a method that can adapt itself as more data is received during inference in an online manner.

Test-time Adaptation (TTA) or Online DA aims to adapt a pre-trained model on-the-fly to the target data without having access to the source dataset. Existing work on TTA [35, 64] focuses on adapting a pre-trained model to a fixed target domain by updating the model's parameters with unsupervised or self-supervised loss. Existing TTA methods are shown to be effective when the target test data is drawn from a stationary domain, but are unstable [46] in a continually changing environment, which is common in real-world applications. Furthermore, as the model has been continu-

Problem Setting	Source Data	Target Data	Train Loss	Test Loss	Open Targets	Forgetting Alleviation
Fine-tuning [59]	✗	x^t, y^t	$\mathcal{L}(x^t, y^t)$	✗	✗	✗
Continual Learning [32]	✗	x^t, y^t	$\mathcal{L}(x^t, y^t)$	✗	✗	✓
Unsupervised DA [73]	x^s, y^s	x^t	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^t, x^s)$	✗	✗	✗
Continual DA [29]	x^s, y^s	x^t	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^t, x^s)$	✗	✗	✓
Source-free DA [22]	✗	x^t	$\mathcal{L}(x^t)$	✗	✗	✗
Domain Generalization [23]	x^s, y^s	✗	$\mathcal{L}(x^s, y^s)$	✗	✓	✗
Test-time training [58]	x^s, y^s	✗	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^s)$	$\mathcal{L}(x^t)$	✓	✗
Test-Time DA (TTA) [64]	✗	✗	✗	$\mathcal{L}(x^t)$	✓	✗
Continual TTA [65]	✗	✗	✗	$\mathcal{L}(x^t)$	✓	✓

Table 1. Characteristics of existing paradigms, the proposed framework DoSe fits in *Continual test-time domain adaptation* setting, which aims to adapt a source pre-trained model to an unseen target domain where the target domain is continually changing. Current work aims to alleviate catastrophic forgetting. DA refers to Domain Adaptation, and TTA refers to Test-Time Domain Adaptation.

ously changed to adapt to new distributions for a long period, knowledge from the source domain becomes more difficult to preserve, resulting in degraded performance known as catastrophic forgetting [13, 34, 43].

Motivated by the Continual Test-time adaptation [65] paradigm, we propose an online continual learning setting for domain incremental semantic segmentation task. The current work adapts any pre-trained model to continually changing target domains without having access to target domain data during training.

Contributions: The key contributions of our work include:

- We propose a novel **Do**main-aware Knowledge **Se**lf-distillation framework that uses domain-aware batch normalization prototypes to improve continual model generalization of urban-scene semantic segmentation task given any pre-trained model.
- To alleviate catastrophic forgetting, we propose an Exemplar-based buffer strategy to store difficult samples in the buffer and replay them after every few iterations to preserve knowledge learned in previously seen domains.
- We show the superiority of our approach over SOTA daytime, UDA, and DG methods for semantic segmentation.
- We conduct extensive experiments, on two real-world datasets ACDC [54], C-Driving [31] and one synthetic dataset SHIFT [47] with six different weather conditions such as overcast, rain, snow, fog, night and cloudy.

2. Related Work

2.1. Adaptation with Target Data

Unsupervised domain adaptation (UDA) methods for semantic segmentation aim to transfer knowledge from labeled source domain to unlabeled target domain [8, 16, 21, 33, 61, 62, 67, 74, 76, 79–81]. However, they assume access to unlabeled target data during training and perform poorly in adverse weather and low illumination conditions. Specialised UDA methods have been designed to perform well in low light conditions such as night [51, 53, 66, 68], and adverse weather conditions such as fog [11, 52]. Still, these methods fail to adapt to continually changing environment.

Continual DA methods [29, 50] aim to continually adapt a model to the unseen domains without degrading performance on previously seen domains. However, existing CDA methods require access to the source dataset, which limits its applicability due to limited storage, and privacy concerns.

Source-Free DA Methods [2, 30, 75] are proposed to relax the necessity of having access to source data. However, they assume that unlabeled target data is available during training, and also they fail to overcome the issue of catastrophic forgetting. Also, it's not practical to assume that it's possible to collect target data during training for all possible low-light conditions possible in the real world.

2.2. Adaptation without Target Data

Domain Generalization (DG) approaches have the objective of increasing generalization on unexplored target data. While DG techniques are popular for image classification [25, 26, 49, 77, 78], very few have been proposed for semantic segmentation. To improve DG in semantic segmentation task, IBNNet [40], Switchable whitening [41], RobustNet [9] and WildNet [23] are proposed. Our proposed approach DoSe improves DG for semantic segmentation and complements the above DG approaches by directly adapting these DG models to continually changing target domains directly during inference.

Test-Time Adaptation (TTA) In Test-Time Training [58], we train a DNN with the main task of classification and an auxiliary task of image rotation prediction. The auxiliary task is used during inference to update model parameters to handle the distribution shift. Fully Test-Time adaptation [64] aims to adapt a source pre-trained model directly to target data during inference. Existing TTA methods are shown to be effective when target data is drawn from a fixed stationary distribution. Also, most of the TTA methods [36, 38, 39, 63, 64, 72] focus only on classification or object detection setting and do not perform well under continuous shifts.

Continual Test-Time Adaptation (CTTA) [65] methods aim to adapt a pre-trained model to a continually changing target environment during inference. Similar to `TTA`, existing works in Continual Test-Time Adaptation [5, 12, 17, 28, 72] focus on classification setting. Recent works CoTTA [65], ONDA [42], and MALL [48] demonstrate improvements in the task of Continual test-time adaptation for urban-scene segmentation. Current work focuses on further enhancing the performance in dealing with catastrophic forgetting of domain-incremental semantic segmentation task.

Batch Normalization (BN) in Domain Shift Batch Normalization statistics vary significantly in case of domain shift between source and target data. Recent work TENT [64], and EATA [38] update the `BN` parameters to align the domain shift between source and target domains. MECTA [17] introduces additional `BN` layers into the network to efficiently adapt in a continual domain adaptation setting. However, it requires access to source data which limits its applicability. Domain-specific `BN` [4] maintains separate batch normalization statistics for source, and target data to handle domain shift. BNE [56] uses domain-specific batch normalization and assigns `BN` statistics of the target domain as a combination of `BN` statistics for multiple source domains in a latent space. Our work enforces the feature matching of `BN` statistics of two batches of images from the same domain to have identical latent representations by distilling knowledge between student and teacher models (referred to as Self-Distillation).

3. Methodology

3.1. Preliminaries

Given a pre-trained model f_{θ_s}, trained on source data $\mathcal{D}_{\mathcal{S}} = \{(x^{\mathcal{S}}, y^{\mathcal{S}})\}$, with network parameters of the source model are denoted by θ_s, Continual Test-Time Adaptation aims to adapt f_{θ_s} to continually changing target domains denoted by $\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_2 ..., \mathcal{X}_T$. Here, $\mathcal{X}_t$ refers to samples from the distribution $\mathcal{P}_{test}$ that continually changes with time t. During inference, we receive a batch of unlabeled images denoted by x_t from $\mathcal{P}_{test}$ which changes w.r.t time denoted by $\mathcal{P}_0, \mathcal{P}_1, ..., \mathcal{P}_t$.

Batch Normalization [20] is a widely used technique for training `DNN` to stabilize the training process and improve convergence speed. Given a Batch size B, channel dimension C, and a feature map $f \in \mathbb{R}^{B \times C \times H \times W}$, the mean and variance of channel i are computed as:

$$\mu_i = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} f_{(b,i,h,w)} \,, \tag{1}$$

$$\sigma_i^2 = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} (f_{(b,i,h,w)} - \mu_c)^2 \,. \tag{2}$$

where μ_i refers to the mean of channel i, and σ_i^2 refers to the variance of channel i. Feature map f is normalized using μ_i, σ_i^2 as follows:

$$BN(f_{(b,i,h,w)}; \mu, \sigma^2) = \gamma_i \frac{f_{(b,i,h,w)} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i \,, \tag{3}$$

where $\epsilon > 0$ ensures the denominator does not get a value 0, and γ, β refer to learnable parameters specific to a layer corresponding to feature map f. μ, σ^2 refer to `BN` mean and variance, respectively.

`BN` layer statistics (mean, variance) are calculated during training and used during inference. However, due to domain shift during inference, global statistics calculated normalize target features incorrectly, leading to significantly poor performance. To overcome this, recent methods [38,64,65] calculate `BN` layer statistics based on the current mini batch's mean and variance. However, due to sudden continual changes in different domains (for example weather), the `BN` mean and variance change rapidly leading to the performance of `BN` to drop rapidly [37].

For every `BN` layer $l \in \{1, 2, ..., L\}$, current mini-batch of target data x_t belonging to domain d, the `BN` prototype b_{x_t} of the current mini-batch is defined as follows,

$$b_{x_t} = \left[b_{x_t}^1, b_{x_t}^2, \ldots, b_{x_t}^L\right] \tag{4}$$

$$= \left[\left(\mu_{x_t}^1, \sigma_{x_t}^{1^2}\right) \oplus \left(\mu_{x_t}^2, \sigma_{x_t}^{2^2}\right) \oplus \ldots \oplus \left(\mu_{x_t}^L, \sigma_{x_t}^{L^2}\right)\right]$$

where $\oplus$ denotes the concatenation operation. We define a domain-aware `BN` prototype P_d and update it as follows:

$$\boldsymbol{P}_d = m\boldsymbol{P}_d + (1-m)\boldsymbol{b}_{x_t} \tag{5}$$

3.2. Self-Distillation Using Batch Norm Prototypes

In the continual adaptation setting, let each iteration is denoted by $k = 1, 2, ...$ Let θ_0 denote the pre-trained model trained on source data, θ_k denote the model updated in the iteration 'k', $\mathcal{L}$ being the loss used to update the model θ_k, and S_k, S_{k-1} denote the probabilities of the segmentation predictions of model θ_k, θ_{k-1}. Here, A(x_t) refers to augmented images of mini-batch x_t. In other words, for each image x in current mini-batch x_t, we perform $b-1$ augmentations, and then, these compiled batches of augmentations of image x along with the image itself x are passed into model θ_{k-1} to get robust predictions for the mini-batch.

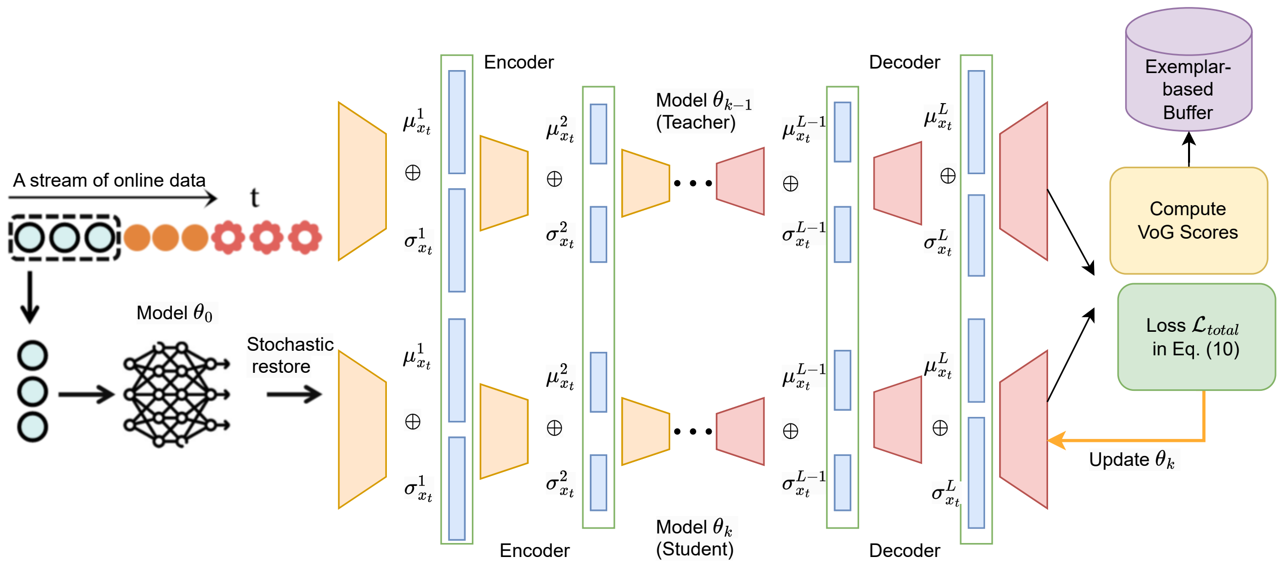

Figure 1. For a batch of images, Our Proposed `DoSe` framework enforces the feature matching between Batch Normalization prototypes between the student (θ_k) and teacher model (θ_{k-1}). Also, the current mini-batch's batch normalization prototype should be closer to the domain-specific batch norm prototype in the Latent space. Exemplar-based replay buffer is used to identify difficult samples based on gradient information and replayed after every few iterations such that model preserves knowledge it learned from previously seen domains. Exemplar-based replay buffer is helpful in alleviating catastrophic forgetting.

The proposed architecture is shown in Fig. 1. We start training the model θ_0 with the target data in the current time stamp denoted by x_t. Note that network θ_{k-1} acts as a teacher to the student model θ_k in self-distillation. We propose a novel self-distillation framework that uses domain-aware BN prototypes to improve continual model adaptation given any pre-trained model during inference.

$$\mathcal{L}_{\text{distill}}^{\text{layer}} = \frac{1}{L} \sum_{l\,\in\{1,\ldots,L\}} \left\| b_{\theta_k}^{l}\left(x_t\right) - b_{\theta_{k-1}}^{l}\left(x_t\right) \right\|_F^2 \tag{6}$$

Further, we define, domain-preserving similarity distillation loss using the BN domain prototype (Eq. (5)) as follows:

$$\mathcal{L}_{\text{distill}}^{\text{domain}} = \frac{1}{L} \sum_{l\in\{1,\ldots,L\}}^{x_t\in d} \left\| b_{\theta_k}^{l}\left(x_t\right) - P_{\theta_{k-1}}^{l}\left(x_t\right) \right\|_F^2 \tag{7}$$

Here $P_{\theta_{k-1}}$ denotes batch domain prototype of domain d.

Motivated by stochastic restoration [65], and to deal with over-fitting and handle catastrophic forgetting. We use the Bernoulli Distribution, to randomly select a subset of weights. p value used to restore a subset of parameters from the source pre-trained model is set to 0.1.

$$M \sim \text{BernoulliDistribution}(p),$$
$$\theta_k = M \odot \theta_0 + (1 - M) \odot \theta_{k-1},$$

Motivated by [60], we use an average of augmented predictions of the image using θ_{k-1} to provide robust pseudo-labels used for training θ_k.

$$\tilde{y}_t = \frac{1}{B} \sum_{i=0}^{B-1} f_{\theta_{k-1}}(\mathrm{A}(x_t)) \tag{8}$$

where $f_{\theta_{k-1}}(\mathrm{A}(x_t))$ corresponds to output of θ_{k-1} with input $\mathrm{A}(x_t)$. We get 1-hot vector label from the teacher model i.e., $\tilde{y}_t$ denoted by y'_t. y'_t is used as a ground truth label to train network θ_k (student model) using cross entropy loss.

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_{i=0}^{B-1} y'_t \log(f_{\theta_k}(x_t)) \tag{9}$$

Total loss is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{distill}^{layer} + \lambda_2 \mathcal{L}_{distill}^{domain} \tag{10}$$

3.3. Exemplar-based Replay buffer

The replay buffer is a well-known technique to alleviate catastrophic forgetting by replaying samples into the neural network so that model does not forget past knowledge. Recent work [45] aims to selectively pick samples in a greedy manner and replay them periodically to alleviate catastrophic forgetting. We extend Variance of Gradients [1] to an online continual setting, where gradients are computed over previous k iteration steps to analyze how gradients vary given a set of input images.

To limit the computational overhead, we use student and teacher model weights, i.e., θ_k and θ_{k-1} to compute the gradients instead of all the models from k iterations. A higher VoG score implies a difficulty level for the model to predict. We maintain a fixed number of samples n per domain in the buffer, and the samples in the buffer keep on updating based on the VoG score in a greedy manner. Images in the Replay buffer are passed into the network after every 100 samples, such that model does not forget the knowledge it learned from previously seen domains. n value is set to 10. The computation of the VoG score is as follows:

$$G = \nabla_{x_t} f_{\theta_k}(x_t), \tag{11}$$

$$\begin{aligned} \mathrm{VoG}_p &= \sqrt{\frac{1}{2}\sum_{t=k-1}^{k}\left(\mathbf{G}_t-\mu\right)^2}, \\ \text{where,} \quad \mu &= \frac{1}{2}\sum_{t=k-1}^{k}\mathbf{G}_t, \end{aligned} \tag{12}$$

$$\text{and} \quad \mathrm{VoG}(\mathrm{x_t}) = \frac{1}{N}\sum_{t=1}^{N}(\mathrm{VoG}_p) \tag{13}$$

where p denotes a pixel in an image, N denotes total number of pixels.

4. Datasets and Evaluation criteria

ACDC [54] ACDC dataset contains 4006 images of adverse weather conditions with a resolution of 1920×1080 pixels. It includes images from foggy, dark, snowy, and rainy conditions. Each visual domain (fog, night, snow, and rain) has 400 training images with ground truth semantic labels, 100 validation images (106 for the night domain), and 500 unlabeled test images.

C-Driving [31] The dataset contains four adverse weather conditions collected from BDD100K dataset [70], consisting of four different weather conditions such as cloudy, rainy, snowy, and overcast conditions. C-Driving rain val results correspond to C-Driving rain validation images.

SHIFT [57] The dataset is a collection of 2.5M synthetic images. We consider the fog, rain, night, and overcast conditions. The SHIFT-val dataset is used to report the results.

Cityscapes (Day) [10] It comprises 5000 broad daylight images of resolution 2048 × 1024 pixels. For validation, Cityscapes-val containing 500 images is used.

Evaluation Criteria Mean Intersection over Union (`mIoU`) is used as an evaluation criterion. Higher `mIoU` indicates better semantic segmentation label predictions. For a fair comparison, multi-scale testing is not considered in all existing and proposed approaches.

Implementation details We implement the proposed approach `DoSe` using Pytorch [44] and train it on an NVIDIA A100 40GB GPU. The network's parameters are updated during training using the Stochastic Gradient Descent (SGD) optimizer with the momentum of 0.9, a learning rate of 1×10^{-3}, and a weight decay of 5×10^{-5}. Hyperparameters for the loss function defined in Eq. (10) (λ_1, λ_2) are 0.1, 1.0 respectively.

Experiment Setup Given a source pre-trained model trained on broad daylight conditions, the objective is to achieve domain incremental continual semantic segmentation, where each domain is a weather condition designed to simulate real-world distribution shifts. The target domain consists of images from various adverse weather and low illumination conditions. To mimic evolving and real-world environments, adverse weather and illumination conditions are revisited to evaluate catastrophic forgetting periodically after every few frames. We repeat a sequence of four weather conditions in 10 iterations (i.e., Fog→Night→Rain→Snow→Fog→Snow, etc...). Each weather is treated as a domain. Existing work in Continual test-time adaptation [12,15,17,28,71,72] focus only on classification setting. For a fair comparison, we extend the classification setting to the segmentation setting and report their results. Recall that recent work CoTTA [65], ONDA [42], TENT [64], MALL [48] focus on the task of test-time adaptation (and not continual learning scenario) for urban-scene segmentation task.

5. Experiments and Results

Comparison with TTA methods As defined in the Problem setup (Sec. 4), using the adverse weather datasets, we demonstrate the results of the domain incremental semantic segmentation task. We report the results of `DoSe` using a source pre-trained model trained on daytime, DeepLabv3+ ResNet101 [7] and RefineNet [27]. On the cityscapes-val dataset, the mIOU of the above-mentioned source pre-trained models is 78.5%, and 71.4%, respectively. The proposed `DoSe` framework significantly improves the performance of DeepLabv3+ ResNet101 by 8% better `mIoU`. On ACDC-snow in the 10th iteration, `DoSe` approach achieves 5.4% better `mIoU` than `CoTTA`. Compared to TENT [64], which does not handle catastrophic forgetting the proposed `DoSe` approach alleviates the catastrophic forgetting by retaining the knowledge it learned in previously seen domains. Average `mIoU` is shown to demonstrate an overall improvement in semantic segmentation performance. Results on RefineNet [27], Qualitative visual results are reported in Fig. 2. Extra qualitative visual results are presented in supplementary material.

Improving SOTA Daytime models We apply the proposed `DoSe` framework on the `SOTA` daytime pre-trained models. We consider `SOTA` daytime pre-trained models namely, Mobilenet V2 [55], LRASPP [18], BiseNetV2 [69], ISANet

Time	t $\longrightarrow$																
Round	1				4				7				10				All
Model+Method	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Mean
DeepLabv3+ ResNet101 [7]	67.6	21.9	52.3	50.6	67.6	21.9	52.3	50.6	67.6	21.9	52.3	50.6	67.6	21.9	52.3	50.6	48.1
• BN Stats Adapt	68.2	20.6	53.5	51.3	64.8	16.5	51.4	48.7	61.3	14.2	48.7	47.4	60.9	13.8	47.9	47.4	44.8
• TENT-continual [64]	67.4	24.7	52.8	48.4	57.5	12.9	40.7	35.6	51.7	11.2	36.1	32.7	48.4	11.3	34.5	31.1	37.3
• EATA-continual [38]	69.7	29.2	56.6	54.8	68.6	27.9	54.3	52.1	64.9	25.7	52.7	51.6	64.9	25.7	52.7	51.6	50.1
• MEMO [72]	67.8	22.3	53.4	51.4	66.7	21.8	52.7	48.4	65.8	20.5	51.5	50.4	48.9	20.7	51.2	50.4	46.5
• SAR [39]	68.5	21.6	55.5	53.3	69.8	22.2	56.7	54.3	69.7	22.8	56.3	54.9	69.7	22.8	56.3	54.9	50.6
• NOTE [15]	69.2	22.4	54.9	54.7	70.4	23.8	55.6	57.3	68.1	20.4	54.2	55.8	68.0	20.2	54.1	55.9	50.3
• RoTTA [71]	69.7	23.8	55.3	55.4	71.8	24.2	54.8	56.9	69.2	21.8	53.3	55.8	67.4	19.8	52.6	54.7	50.4
• MECTA [17]	69.4	23.1	54.8	54.9	70.1	24.4	55.4	55.1	70.5	24.1	55.9	56.1	69.3	22.4	54.9	54.7	50.8
• RATP [28]	70.3	21.5	54.6	52.1	69.4	21.3	54.8	54.2	69.3	21.5	55.1	54.6	68.8	20.7	54.8	54.2	49.9
• RMT [12]	69.9	26.2	56.9	54.3	70.4	26.6	57.1	55.3	66.9	21.3	54.1	50.8	65.4	21.1	53.3	50.4	50.1
• MALL [48]	70.4	25.6	56.4	52.1	71.8	26.6	57.7	53.2	66.7	23.1	53.4	52.7	63.4	20.2	51.8	51.4	49.8
• ONDA [42]	71.5	25.1	57.2	52.4	72.1	27.3	56.6	54.4	67.2	24.8	52.2	53.3	64.8	20.8	52.7	52.6	51.4
• CoTTA [65]	71.6	24.5	56.9	56.2	71.1	25.4	59.7	57.7	72.6	24.7	59.5	57.8	73.2	23.8	57.9	55.3	53.2
• `DoSe` **(ours)**	**73.4**	**26.4**	**57.7**	**57.9**	**76.5**	**27.9**	**60.4**	**60.3**	**76.1**	**28.2**	**61.1**	**60.4**	**76.8**	**30.9**	**63.8**	**60.7**	**56.1**

Table 2. Comparison of `DoSe` with existing TTA methods on DeepLabv3+ ResNet101 [7] pre-trained model. Results are reported after applying existing TTA methods on the DeepLabv3+ ResNet101 [7] pre-trained model trained on source (daytime) data. Results of our `DoSe` framework are reported in pink colored row. The experiment setup is described in Sec. 4.

(a) Input image (b) RATP (c) CoTTA (d) `DoSe` (ours) (e) Semantic GT

Figure 2. Qualitative visual comparison of our proposed `DoSe` framework with existing state-of-the-art `TTA` methods.

[19], STDC [14], and GCNet [3]. Results are reported in Tab. 3. We observe that `DoSe` improves the performance significantly across multiple adverse weather conditions. `DoSe` improves BiseNet V2 [69] average miou performance by 6%. The performance improvement of `DoSe` framework across multiple `SOTA` daytime models demonstrate the generalization ability of the proposed framework.

Improving SOTA Domain Generalization methods In Tab. 4, we demonstrate the effectiveness of the `DoSe` framework over the `SOTA` `DG` methods for urban-scene segmentation WildNet [23], RobustNet [9], IBNNet [40], and Switchable whitening (SW) [41]. We apply the proposed `DoSe` framework to the pre-trained models. `DoSe` on IBN-Net, WildNet shows an average increase of 2.6%, 3.1% better `mIoU` respectively across 10 iterations. This illustrates that `DoSe` can further enhance the performance of `DG` approaches to adapt to unseen adverse weather images without altering training or requiring unlabeled target data during training.

Improving SOTA UDA methods Most `UDA` methods for semantic segmentation focus on adverse weather and low illumination conditions as there is limited availability of ground truth data in low light conditions. We demonstrate performance improvement on `SOTA` `UDA` methods designed for night image segmentation in two settings: 1) No availability of night images during training, 2) Access to unlabeled night images during training. For setting 1, we consider ZeroShotDN [24]. For setting 2, we consider MGCDA [53], DANNet [66] as our pre-trained model and adapt them to continually changing target domains. Results are reported in Tab. 7. `DoSe` enhances the performance of `SOTA` method for night image segmentation by 2.7% better `mIoU`.

6. Ablation Study

To demonstrate the effectiveness of **Do**main-aware **Se**lf-Distillation method (Sec. 3.2) and Exemplar-based buffer (Sec. 3.3), we perform an ablation study with and without these loss components in the Total loss function defined in Eq. (10). Results are reported in Tab. 6.

Results on SHIFT, C-driving datasets Results of the proposed `DoSe` framework on SHIFT, C-Driving are reported in Tab. 8 and Tab. 9 respectively. Results show that `DoSe` is effective in improving performance significantly on SHIFT,

Time	t ⟶																
Round	1				4				7				10				All
Model+Method	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Mean
Mobilenet V2 [55]	41.7	4.4	38.3	24.1	41.7	4.4	38.3	24.1	41.7	4.4	38.3	24.1	41.7	4.4	38.3	24.1	27.1
• DoSe (ours)	**42.3**	**7.8**	**40.5**	**27.2**	**44.6**	**10.1**	**41.1**	**28.4**	**45.1**	**11.5**	**42.3**	**29.8**	**45.9**	**11.8**	**42.5**	**30.3**	**31.3**
LRASPP [18]	51.2	15.4	37.3	38.2	51.2	15.4	37.3	38.2	51.2	15.4	37.3	38.2	51.2	15.4	37.3	38.2	35.5
• DoSe (ours)	**52.3**	**16.8**	**39.8**	**40.1**	**54.6**	**17.5**	**40.9**	**41.3**	**55.2**	**18.2**	**41.5**	**42.8**	**56.7**	**20.3**	**43.2**	**43.3**	**39.1**
BiseNetV2 [69]	48.3	13.7	38.2	35.7	48.3	13.7	38.2	35.7	48.3	13.7	38.2	35.7	48.3	13.7	38.2	35.7	33.9
• DoSe (ours)	**50.1**	**16.4**	**39.4**	**36.8**	**51.8**	**18.2**	**40.9**	**40.2**	**53.4**	**20.1**	**44.3**	**43.4**	**55.8**	**20.9**	**46.9**	**44.2**	**38.9**
ISANet [19]	62.2	18.8	46.8	44.9	62.2	18.8	46.8	44.9	62.2	18.8	46.8	44.9	62.2	18.8	46.8	44.9	43.1
• DoSe (ours)	**63.1**	**20.4**	**48.2**	**46.5**	**64.5**	**21.6**	**51.3**	**49.2**	**65.3**	**23.3**	**54.9**	**50.5**	**67.5**	**23.9**	**55.2**	**50.8**	**47.3**
STDC [14]	62.6	18.7	46.4	45.3	62.6	18.7	46.4	45.3	62.6	18.7	46.4	45.3	62.6	18.7	46.4	45.3	43.2
• DoSe (ours)	**64.3**	**20.4**	**49.8**	**46.7**	**64.9**	**21.7**	**52.3**	**48.9**	**65.4**	**22.5**	**54.7**	**51.5**	**65.9**	**23.4**	**56.4**	**52.8**	**47.6**
GCNet [3]	62.6	19.4	47.8	48.1	62.6	19.4	47.8	48.1	62.6	19.4	47.8	48.1	62.6	19.4	47.8	48.1	44.5
• DoSe (ours)	**63.3**	**20.2**	**48.7**	**50.4**	**65.8**	**21.1**	**50.2**	**50.8**	**66.4**	**21.4**	**51.1**	**51.4**	**67.9**	**21.8**	**52.4**	**51.6**	**47.1**

Table 3. Performance of various SOTA techniques with and without using our DoSe framework. Results after using our DoSe framework are reported in pink colored row. Experiment setup is described in Sec. 4.

Time	t ⟶																
Round	1				4				7				10				All
Model+Method	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Mean
IBNNet [40]	63.6	25.9	50.4	50.2	63.6	25.9	50.4	50.2	63.6	25.9	50.4	50.2	63.6	25.9	50.4	50.2	47.5
• DoSe (ours)	**64.5**	**26.2**	**51.8**	**52.4**	**66.7**	**27.4**	**53.8**	**55.4**	**67.4**	**28.8**	**55.3**	**56.2**	**67.8**	**30.2**	**55.8**	**57.3**	**51.1**
SW [41]	54.8	20.7	40.2	45.6	54.8	20.7	40.2	45.6	54.8	20.7	40.2	45.6	54.8	20.7	40.2	45.6	40.3
• DoSe (ours)	**56.4**	**21.6**	**42.8**	**47.1**	**57.7**	**22.9**	**43.4**	**47.5**	**59.9**	**24.3**	**45.4**	**49.8**	**60.1**	**25.2**	**46.5**	**50.3**	**43.1**
RobustNet-R50 [9]	60.5	26.1	46.9	50.1	60.5	26.1	46.9	50.1	60.5	26.1	46.9	50.1	60.5	26.1	46.9	50.1	45.9
• DoSe (ours)	**62.1**	**27.6**	**47.7**	**50.8**	**63.2**	**28.8**	**49.9**	**52.2**	**64.4**	**30.2**	**52.1**	**52.7**	**64.8**	**30.5**	**52.3**	**53.3**	**48.9**
RobustNet-R101 [9]	59.3	26.8	46.5	45.2	59.3	26.8	46.5	45.2	59.3	26.8	46.5	45.2	59.3	26.8	46.5	45.2	44.4
• DoSe (ours)	**60.1**	**27.7**	**48.9**	**46.4**	**61.4**	**28.5**	**49.4**	**47.3**	**62.1**	**29.2**	**50.8**	**48.2**	**63.7**	**30.4**	**51.1**	**50.8**	**47.2**
WildNet [23]	64.8	27.9	50.5	49.4	64.8	27.9	50.5	49.4	64.8	27.9	50.5	49.4	64.8	27.9	50.5	49.4	48.1
• DoSe (ours)	**66.4**	**28.2**	**51.8**	**49.9**	**67.9**	**29.9**	**54.4**	**50.5**	**69.4**	**30.4**	**55.6**	**51.5**	**70.6**	**31.3**	**56.4**	**53.9**	**51.2**

Table 4. Results of DoSe framework on SOTA pre-trained Domain Generalization models. Results are reported after applying DoSe framework on SOTA DG methods. The experiment setup is described in Sec. 4. RobustNet-R50 corresponds to RobustNet-ResNet50, and RobustNet-R101 corresponds to RobustNet-ResNet101.

Method	**Inference time**
DeepLabv3+ ResNet101 [7]	80
CoTTA [65]	920
DoSe (ours)	750

Table 5. Inference time comparison of DoSe framework on DeepLabv3+ ResNet101, Inference time values are in seconds.

C-Driving datasets.

Affect of order sequence To understand how the sequence order of domains affect the performance of the proposed DoSe framework, we vary the order of domains from the sequence used to report performance in Tab. 2. Results are reported in supplementary material.

Computational cost In order to analyze the computational cost of the proposed DoSe framework, we consider the DeepLabv3+ ResNet101 [7] and adapt it to 100-night images from ACDC [54] and perform continual test-time adaptation for 10 iterations. Results are reported in Tab. 5. Proposed DoSe framework adapts to continually changing target environment faster than existing SOTA method CoTTA.

Time	t →																
Round	1				4				7				10				All
Model+Method	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Mean
DeepLabv3+ MobileNet V2 [7]	41.8	4.4	38.3	24.1	41.8	4.4	38.3	24.1	41.8	4.4	38.3	24.1	41.8	4.4	38.3	24.1	27.1
• **DoSe (ours)**	**52.2**	**7.5**	**44.6**	**29.2**	**54.2**	**10.8**	**47.7**	**35.7**	**56.8**	**13.2**	**45.4**	**38.6**	**57.2**	**11.4**	**45.9**	**37.2**	**36.8**
• w/o Self-Distillation	46.5	5.1	43.5	28.8	51.3	7.4	44.2	32.9	53.7	10	43.9	35	53.4	10.9	42.3	34.6	32.3
• w/o Exemplar-based Buffer	49.3	6.5	44.7	29.8	50.4	8.2	45.8	33.3	52.8	9.7	45.9	35.4	54.1	9.4	44.9	35.3	34.7

Table 6. To analyse the impact of Self-distillation method and Exemplar-based replay buffer, we perform ablation study of the proposed framework DoSe using different loss variants on DeepLabv3+ MobileNet V2 [7]. The experiment setup is described in Sec. 4.

Time	t →																
Round	1				4				7				10				All
Model+Method	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Fog	Night	rain	snow	Mean
Trained on source data (daytime images) only																	
ZeroShotDN [24]	42.4	32.8	28.9	31.7	42.4	32.8	28.9	31.7	42.4	32.8	28.9	31.7	42.4	32.8	28.9	31.7	33.9
• **DoSe (ours)**	**43.1**	**33.5**	**30.4**	**32.8**	**44.2**	**34.1**	**31.1**	**33.2**	**44.7**	**34.6**	**31.4**	**35.4**	**45.9**	**35.4**	**33.1**	**37.8**	**36.2**
Trained on source data (daytime images) and unlabeled (nighttime) target images only																	
MGCDA [53]	49.4	38.6	35.8	44.9	49.4	38.6	35.8	44.9	49.4	38.6	35.8	44.9	49.4	38.6	35.8	44.9	42.1
• **DoSe (ours)**	**51.2**	**39.4**	**36.4**	**45.7**	**52.8**	**40.1**	**37.3**	**46.3**	**53.4**	**41.3**	**38.1**	**47.4**	**54.1**	**42.5**	**38.7**	**47.8**	**44.5**
DANNet [66]	52.4	39.9	45.7	49.1	52.4	39.9	45.7	49.1	52.4	39.9	45.7	49.1	52.4	39.9	45.7	49.1	46.7
• **DoSe (ours)**	**53.1**	**40.4**	**46.5**	**50.5**	**54.5**	**41.1**	**48.2**	**50.8**	**55.1**	**41.5**	**50.7**	**51.7**	**56.8**	**42.3**	**52.9**	**53.4**	**49.4**

Table 7. Results of DoSe framework on SOTA UDA methods for semantic segmentation. The experiment setup is described in Sec. 4.

Time	t →																
Round	1				4				7				10				All
Model+Method	Fog	Night	rain	overcast	Fog	Night	rain	overcast	Fog	Night	rain	overcast	Fog	Night	rain	overcast	Mean
DeepLabv3+ ResNet101 [7]	28.5	21.8	24.3	29.2	28.5	21.8	24.3	29.2	28.5	21.8	24.3	29.2	28.5	21.8	24.3	29.2	25.9
• **DoSe (ours)**	**29.9**	**24.1**	**28.2**	**32.3**	**31.3**	**25.7**	**30.6**	**32.7**	**31.1**	**25.3**	**30.5**	**32.5**	**30.4**	**25**	**30.3**	**32.2**	**29.5**

Table 8. Results of DoSe framework using DeepLabv3+ ResNet101 [7] on SHIFT dataset. The experiment setup is described in Sec. 4.

Time	t →																
Round	1				4				7				10				All
Model+Method	cloud	rain	snow	overcast	cloud	rain	snow	overcast	cloud	rain	snow	overcast	cloud	rain	snow	overcast	Mean
DeepLabv3+ ResNet101 [7]	42.3	34.8	35.5	44.7	42.3	34.8	35.5	44.7	42.3	34.8	35.5	44.7	42.3	34.8	35.5	44.7	39.3
• **DoSe (ours)**	**43.9**	**35.3**	**36.7**	**46.9**	**43.9**	**35.3**	**36.7**	**47.3**	**44.5**	**36.3**	**38.2**	**47.5**	**45.6**	**38.9**	**39.9**	**48.1**	**41.6**

Table 9. Results of DoSe framework using DeepLabv3+ ResNet101 [7] on C-driving dataset. The experiment setup is described in Sec. 4.

7. Conclusion

Existing test time adaptation methods fail to perform well due to the assumption that samples come from a single target distribution. However, in real-world applications, the target domain keeps changing continually with time. To alleviate this, we introduce a novel **Do**main-aware **Se**lf-distillation method to enforce feature matching in the latent space between batch normalization prototypes of student and teacher models. To alleviate catastrophic forgetting, we propose an Exemplar-based replay buffer to identify difficult samples for the model to predict and replay them to preserve the learned knowledge from previously seen domains. Proposed DoSe framework outperforms SOTA methods in various settings, viz: naive pre-training with daytime samples, domain generalization, and unsupervised domain adaptation.

Acknowledgment: We thank UQ-IIT Delhi Research Academy (UQIDAR) for providing a contingency grant. This work has also been partly supported by the funding received from DST through the IMPRINT program (IMP/2019/000250).

References

[1] Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2022. 4

[2] Waqar Ahmed, Pietro Morerio, and Vittorio Murino. Continual source-free unsupervised domain adaptation. *arXiv preprint arXiv:2304.07374*, 2023. 2

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Global context networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 6, 7

[4] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. 3

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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. 1

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 5, 6, 7, 8

[8] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018. 2

[9] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 2, 6, 7

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[11] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128(5):1182–1204, May 2020. 2

[12] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 3, 5, 6

[13] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. In *ECCV*. Springer, 2020. 2

[14] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021. 6, 7

[15] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5, 6

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 2

[17] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. Mecta: Memory-economic continual test-time model adaptation. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5, 6

[18] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1314–1324, October 2019. 5, 7

[19] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019. 6, 7

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3

[21] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12975–12984, 2020. 2

[22] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R. Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7046–7056, October 2021. 2

[23] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 2, 6, 7

[24] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 6, 8

[25] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021. 2

[26] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-source style augmentation for improved domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 509–519, 2023. 2

[27] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 5

[28] Chenxi Liu, Lixu Wang, Lingjuan Lyu, Chen Sun, Xiao Wang, and Qi Zhu. Deja vu: Continual model generalization for unseen domains. *arXiv preprint arXiv:2301.10418*, 2023. 3, 5, 6

[29] Hong Liu, Mingsheng Long, Jianmin Wang, and Yu Wang. Learning to adapt to evolving domains. *Advances in neural information processing systems*, 33:22338–22348, 2020. 1, 2

[30] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2

[31] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[32] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 2

[33] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2

[34] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[35] Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021. 1

[36] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 3

[37] Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *CoRR*, abs/2006.10963, 2020. 3

[38] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The International Conference on Machine Learning*, 2022. 3, 6

[39] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023. 3, 6

[40] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2, 6, 7

[41] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1863–1871, 2019. 2, 6, 7

[42] Theodoros Panagiotakopoulos, Pier Luigi Dovesi, Linus Härenstam-Nielsen, and Matteo Poggi. Online domain adaptation for semantic segmentation in ever-changing conditions. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 5, 6

[43] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 2

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

[45] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 524–540. Springer, 2020. 4

[46] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *ICCV*, pages 8558–8567, 2021. 1

[47] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009. 2

[48] Nikhil Reddy, Abhinav Singhal, Abhishek Kumar, Mahsa Baktashmotlagh, and Chetan Arora. Master of all: Simultaneous generalization of urban-scene segmentation to all adverse weather conditions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 51–69. Springer, 2022. 3, 5, 6

[49] Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *arXiv preprint arXiv:2102.11436*, 2021. 2

[50] Mohammad Rostami. Lifelong domain adaptation via consolidated internal distribution. *Advances in neural information processing systems*, 34:11172–11183, 2021. 1, 2

[51] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019. 2

[52] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *European Conference on Computer Vision (ECCV)*, pages 707–724, 2018. 2

[53] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 6, 8

[54] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. *arXiv preprint arXiv:2104.13395*, 2021. 2, 5, 7

[55] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5, 7

[56] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115, 2023. 3

[57] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 5

[58] Yu Sun, Xiaolong Wang, Liu Zhuang, John Miller, Moritz Hardt, and Alexei A. Efros. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 2

[59] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 2

[60] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeuIPS*, pages 1195–1204, 2017. 4

[61] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2

[62] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 2

[63] Vibashan VS, Poojan Oza, and Vishal M Patel. Towards online domain adaptive object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 478–488, 2023. 3

[64] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2, 3, 5, 6

[65] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 2, 3, 4, 5, 6, 7

[66] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 2, 6, 8

[67] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018. 2

[68] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. Cdada: A curriculum domain adaptation for nighttime semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2962–2971, 2021. 2

[69] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 5, 6, 7

[70] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[71] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. 5, 6

[72] M. Zhang, S. Levine, and C. Finn. MEMO: Test time robustness via adaptation and augmentation. 2021. 3, 5, 6

[73] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. *arXiv preprint arXiv:2101.10979*, 2021. 2

[74] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018. 2

[75] Yuyang Zhao, Zhun Zhong, Zhiming Luo, Gim Hee Lee, and Nicu Sebe. Source-free open compound domain adaptation in semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7019–7032, 2022. 2

[76] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 2

[77] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 2

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

[79] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2

[80] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2

[81] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 2