Multi-Domain Conditional Image Translation: Translating Driving Datasets from Clear-Weather to Adverse Conditions

Vishal Vinod, K. Ram Prabhakar, R. Venkatesh Babu, Anirban Chakraborty
Department of Computational and Data Sciences,
Indian Institute of Science Bangalore, India

Abstract

Vision systems for fully autonomous navigation must perform well even in unstructured and degraded scenarios. In most driving datasets today, there is a bias toward clear-weather conditions as compared with extreme-weather owing to the difficulty in capturing and annotating large-scale image datasets degraded by adverse weather. While there has been extensive research on techniques such as deraining, dehazing and on tasks such as segmentation and domain adaptation, there has been minimal attention toward methods to effectively translate clear-weather driving datasets to extreme-weather domains. To address this, we present a method that builds on recent advances in Generative Networks and Self-Supervised Learning to perform conditional multi-domain image translation. We evaluate our method on the semantic scene understanding task and demonstrate quantitatively superior translation results from clear-weather conditions to adverse-weather shifted domains such as Rain, Night and Fog conditions. From our experiments, we show improved domain invariant content disentanglement, and segmentation methods trained with datasets translated using the proposed method have improved performance over single and multi-domain image translation baselines on real-world adverse weather data.

1. Introduction

The performance of vision algorithms are crucial for applications in autonomous navigation and robotic systems that are to be deployed in real-world settings. This invariably involves unstructured environments and degraded scenes from non-ideal weather and illumination settings, and the vision algorithms must be able to perform optimally even in such adverse conditions. Image degradation due to fog [65], rain [33, 71] and low-illumination from night-time scenes [6, 16] lead to a significant decrease in visibility [80] and that leads to decreased performance across vision tasks such as segmentation and detection. An autonomous system is required to be robust and perform well even in degraded conditions such as fog, rain, night-time glare, overcast conditions and snow. While a significant body of research has dealt with deraining [28, 33, 62, 71], defogging/dehazing [52, 73] and night-to-day translation [61, 83], novel techniques for the multi-domain visually degraded scene generation problem to augment existing driving datasets captured in clear-weather to adverse weather conditions has received minimal attention. Compiling large-scale driving datasets in degraded scenarios such as Rain, Fog, Snow and Night-time conditions is not scalable due to unpredictable weather patterns (while capturing the data) and due to poor illumination and scene degradation that make the task of annotating these datasets a formidable challenge. As compared with clear-weather daytime driving datasets such as Cityscapes [14], KITTI [19] and IDD [70], degraded datasets are significantly smaller [58]. While the BDD100K dataset [78] has approximately 40% of its images captured at night, only 345 images have semantic segmentation labels [60]. Further, nearly 70% of the labels for BDD100K Night images have labeling errors [57]. Hence, there is a need for a multi-domain generative model to effectively translate large clear-weather driving datasets to adverse weather domains with realistic results. Further, dataset translation also allows the reuse of existing high-quality annotations in addition to augmenting and enriching existing datasets.

CNNs are heavily data-sensitive. In Table 1, the semantic segmentation performance of a BiSeNetV2 [77] model trained on the Cityscapes dataset [14] shows a significant decrease in mIoU when evaluated on real-world datasets such as Foggy Zürich [15, 58], ACDC-Rain [60] and Dark Zürich [16] for fog, rain and night respectively. Deep vision models trained for vision and perception tasks in autonomous navigation show an alarming bias toward clear-weather daylight conditions owing to the model architectures being trained on largely available annotated data in the well illuminated domain. This dataset bias leads to a lack of generalization to adverse weather conditions. The research community will benefit from an improved multi-domain

Table 1. Comparison between semantic segmentation models trained on Cityscapes and evaluated on adverse-weather domain datasets (in Mean IoU). The models used are: BiSeNetv2 [77], PSPNet [82], DeepLabv3+ [8], DDRNet [27].

Semantic Segmentation Models Trained on Cityscapes (mIoU)				
Validation Data (↓)	BiSeNetv2	PSPNet	DeepLabv3+	DDRNet
Cityscapes	75.75	76.54	77.78	79.63
Foggy Zürich	11.22	15.35	19.40	32.55
Foggy Driving	22.50	28.39	22.79	40.85
Foggy-Cityscapes	39.05	51.82	56.23	62.47
ACDC-Rain	22.05	39.32	29.11	41.37
Rainy-Cityscapes	17.65	41.28	46.05	45.73
Dark Zürich	1.43	8.17	7.02	14.61

translation method that can be used to obtain a domain-shifted dataset with realistic images. Such a method can be used to translate clear-weather datasets such as Cityscapes [14] or the Indian Driving Dataset [70] to domain-shifted versions in Fog, Rain or Night-time conditions. In this work, we utilize images from several datasets (refer Section 3) spanning the domains of Rain, Fog, and Night-time scenes to train the multi-domain image translation method.

Recently, several works such as DSMAP [5], MSGAN [44] and DRIT [37] have investigated methods to enable diverse image translation by improving the disentanglement of domain-invariant and domain-specific content spaces to translate an image from a source domain to diverse time-of-day or weather target domains. While these methods and the multi-domain image translation methods such as MUNIT [31] and DRIT++ [38] show good performance for diverse image synthesis, their translation results when used for semantic scene understanding tasks in adverse weather such as SFSU [58] show suboptimal generalization performance. Moreover, our experiments with high-resolution DRIT++ with a modified perceptual loss [83] and mode-seeking regularizer [44] demonstrated satisfactory diverse image synthesis only for similar scenic datasets such as Cityscapes and BDD100K whereas the translation performance for the Indian Driving Dataset (IDD) dataset had several artifacts (IDD has a wide variety of scenes with more segmentation label classes than Cityscapes). We note for these methods that despite the use of several loss functions to improve disentanglement of domain-invariant and domain-specific content spaces, their translation of style across datasets with scenic domain-shift lead to issues in translational generalization. Our method addresses this issue by using a contrastive loss [50], multi-scale normalization [30] and denormalization [51] methods to effectively preserve the content information and apply style in a multi-scale manner leading to better generalization and translation performance.

We address these challenges by designing a novel multi-domain image translation approach drawing on insights from ForkGAN [34], TSIT [34] and FastCUT [50] for our method. In Fig.(1), we show the pipeline of our

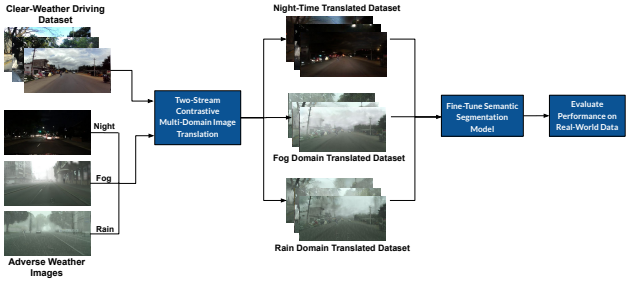


Figure 1. Overview of the multi-domain translation method.

method trained with a clear-weather dataset conditioned on adverse-weather images to translate the dataset to an adverse domain-shifted version. Experiments show that our method outperforms prior image translation methods for the semantic scene understanding task in adverse conditions and demonstrates improved generalization capability when a semantic segmentation model is fine-tuned with a domain-shifted dataset translated using our method. In summary, the contributions of this work are:

- We propose a two-stream contrastive multi-domain image translation method with Adaptive Instance Normalization [30] to introduce style and Adaptive Denormalization to preserve content [34, 51].
- We show that using a multi-scale contrastive loss [10, 45, 48, 50] and a perceptual loss, we are able to better preserve the content and improve the style infusion from training with images from Foggy-Cityscapes (fog) and Rainy-Cityscapes (fog + rain) datasets.
- We show that with less than 170 images from each target domain, the proposed approach can match the performance of baseline multi-domain and single-domain translation methods on the semantic scene understanding evaluation on real-world adverse weather datasets.
- We present experiments and ablation studies to illustrate the effectiveness of the proposed approach.

2. Related Work

Multi-domain image translation methods to transform datasets from one domain to many has received less attention from the research community. There are several datasets that have been compiled, yet most driving datasets are pertinent to clear-weather conditions. The focus of the research community has been on only a select few clear-weather datasets for evaluation on segmentation and detection tasks. We discuss related works and experiments:

Multi-Domain Image Translation. Image-to-image translation methods such as CycleGAN [84], MUNIT [31], FUNIT [41] have successfully been able to translate images from one domain to another in an unpaired setting. Some methods [1, 13, 30, 37, 38] have performed well in multi-

domain translation and diverse translation tasks [37, 40], with good performance even in high resolution [66]. Image augmentation is an important step in self-supervised learning [10, 24], and more recent methods in Domain Adaptation [39, 49, 68, 69] use these image translation [12] results for the synthetic to real (syn2real) task. Further, recent methods such as COCO-FUNIT [56] and unsupervised methods such as [2] have been successful in the low-data regime but their performance has not been investigated in the clear-weather to adverse weather domains. Still, most methods require large datasets in the target domain - which is sometimes infeasible or not scalable to compile (Sec.(1)).

Style Transfer. Style transfer methods preserve the content information and infuse style to manipulate the image [7, 18]. Recent works propose GAN-based methods [31, 79, 84], wavelet transforms [76] or graph cuts [81] for effective style transfer. Further, [43, 74] propose universal photo-realistic style transfer. To better preserve the scene information in the target domain, we experimented with improving multi-scale adaptive normalization [30] for learning a joint embedding onto a bilateral grid based on [74]. This enables edge-aware local affine transforms for photo-realistic multi-domain translation. This approach transforms the generative adversarial formulation into an edge-aware style transfer method that suitably combines local and global scene information across domains. We found better scene consistency for the IDD dataset (refer Sec.(1)) with much clearer and sharper translations for '*two-wheelers*' and '*autorickshaws*' that are absent in weather-shifted domain scenes. While the translation quality and the photo-realism of the generated images were qualitatively good, the intensity of the weather degradation was low and had more global than local style in the output leading to mild translations.

Physics-based degradation. Recently, Physics-based methods have been proposed to approximate and render fog [15, 58] and rain [23, 28, 33, 67, 71, 75] with realistic photometry and physical properties. While the rendering is realistic, these methods depend on availability of depth information [28, 58], camera calibration [15], inputs where pixel is not covered by rain [71], streaks radiance [23] etc. Moreover, these methods include computations such as depth-completion, guided-filtering, per-pixel rain masks, simulations etc., leading to long render-times for each image.

Domain Adaptation. Few-shot learning [17, 63, 64] and Domain Adaptation [4, 29, 53] methods have been studied in order to adapt to unknown domains in an unsupervised manner. Recent works like [55] have described the concept of Universal Domain Adaptation (UniDA) to be able to adapt to any target domain with arbitrary shifts. In addition, techniques in realistic settings such as Open-Compound Domain Adaptation (OCDA) [42, 49] have been proposed to adapt a model trained on clear-weather driving datasets with labels (Indian Driving Dataset) to achieve better performance on multiple heterogeneous weather domains. Improving methods for realistic DA like OCDA can be an important step toward UniDA. A few-shot Domain Adaptation formulation [47] has been proposed and recent work [54] using few labeled samples and many unlabeled samples in the target domain have been studied for classification. Curriculum learning methods [57, 59] have proposed techniques to improve segmentation performance in adverse weather conditions. Thus, there are few works that address the adaptation to adverse weather conditions and there is a need for datasets to have adverse domain-shifts with the same labels to evaluate adaptation performance.

3. Methodology

3.1. Problem Formulation

The two-stream contrastive multi-domain image translation model is illustrated in Fig.(2) with the source stream and style conditioning stream [34]. We denote the source domain ($\mathbf{S}$) comprising of clear-weather images $\{S_n\}$ as the source input. For this work, we consider three adverse condition domains as the target styles - Fog, Rain and Nighttime; Our setting is unpaired, hence we use the images from the adverse target domains ($\{T_{fog}, T_{rain}, T_{night}\} \in \mathbf{T}$) as our style conditioning input, where $\mathbf{T}$ is the combined composite style conditioning domain. The $\mathbf{S}$ is the source content input and $\mathbf{T}$ is the target style input. We train the two-stream conditional CNN model ($\mathbb{N}$) to capture the source domain content structure and the composite target domain styles at multiple feature scales of the upsampling block feature maps. The model $\mathbb{N}$ consists of a source stream that takes as input $\widehat{\mathbf{S}} = \widehat{S}_n, \forall n = 1, \cdots, N$ and a style conditioning stream that takes in $\widehat{\mathbf{T}} = \widehat{T}_m, \forall m = 1, \cdots, M$ as input. The Generator $\mathbf{G}$ samples a latent vector $z_0 \in \mathbf{Z}$ from a random Gaussian distribution and progressively utilizes the target style and source content captured by the AdaIn and FADE blocks respectively in the feature maps of the i-th residual block (of k total blocks). In addition to progressively refining the generated image at various feature scales, we introduce a contrastive loss to maximize the mutual information [48] between the source input and the generated image to improve the translation results. Further, multi-scale discriminators [72] enable the effective infusion of adverse-weather domain style in the generated images.

3.2. Proposed Approach

AdaIn. We use the Adaptive Instance Normalization [30] for the style image feature maps at each of the i-th blocks (i=7) in the style conditioning stream. This enables better preservation of the style at various scales to fuse the information in a multi-scale manner with the source content features captured by the denormalization block (FADE). We denote the feature map at the i-th block to be m_s^i, the fea-

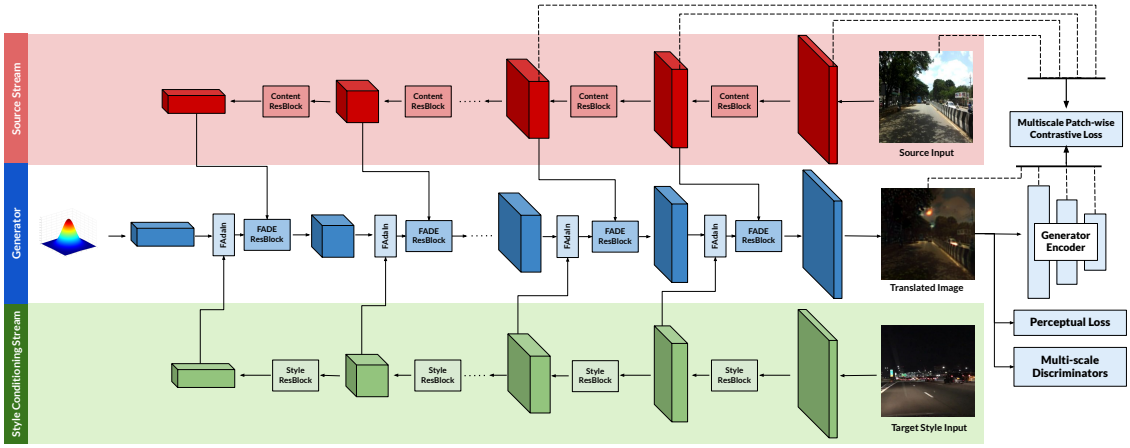


Figure 2. Overview of the two-stream contrastive multi-domain translation model.

tures extracted (AdaIn) to be f_s^i, and the μ and σ represent the mean and standard deviation respectively. The AdaIn formulation is as follows:

$$AdaIn\ (m_s^i, f_s^i) = \sigma(f_s^i)\left(\frac{m_s^i - \mu(m_s^i)}{\sigma(m_s^i)}\right) + \mu(f_s^i) \quad (1)$$

FADE. To fuse the domain-invariant information captured by the source stream, we use the denormalization method described by [34, 51] for the source feature maps at each of the i-th blocks (i=7) in the source stream. First, the feature map m_i is subject to batch normalization [32] followed by the learned denormalization. The denormalization parameters modulate the scale and offset from batch normalization by using one-layer convolutions to learn a scale factor γ_c^i and and offset β_c^i. Using the same notation as above, m_c^i is the feature map at the i-th block, f_c^i are the features extracted by the FADE block, and μ and σ represent the mean and standard deviation respectively. The FADE formulation is as follows:

$$FADE\ (m_c^i, f_c^i) = \gamma_c^i \cdot \frac{m_c^i - \mu(m_c^i)}{\sigma(m_c^i)} + \beta_c^i \quad (2)$$

As discussed, our method aims to translate images from the clear-weather domain conditioned on images in the weather domain (style) to generate realistic images in the adverse weather domain with the aim of translating the entire dataset to the domain-shifted version. We use the AdaIn [30] to preserve style information and FADE [34, 51] to preserve domain-invariant content information for each feature map of the i-th residual block as shown in Fig.(2).

Method Overview. Referring to Fig.(1) and Fig.(2), we describe the training flow and lay the context for the proposed method. An image $\{S_n\}$ from the clear-weather source domain $\mathbf{S}$ is the content stream input. Similarly, an image $\{T_m\}$ from the adverse-weather composite (Fog, Rain and Night-time) target domain ($\{T_{fog}, T_{rain}, T_{night}\} \in \mathbf{T}$), is the style stream conditioning input. The generator $\mathbf{G}$ draws a latent vector z_0 from a random Gaussian distribution and utilizes, (a) the style feature map m_{style}^i and style features f_{style}^i to fuse the target domain style with the content, and (b) the content feature map $m_{content}^i$ and style features $f_{content}^i$ to fuse the source domain content with the style at each of the i-th blocks ($i = 7$) to perform multi-domain image-to-image translation. The output from the generator is the translated image in the conditioned domain. We compute the adversarial loss ($\mathcal{L}_{GAN}$) with a multi-scale discriminator to improve the realism of the translation, and compute the modified perceptual loss ($\mathcal{L}_{Perceptual}$) for improved content consistency. A multi-scale patch-wise contrastive loss ($\mathcal{L}_{CTR}$) is used to preserve source domain content and also translate the image with a less restrictive assumption than methods that impose the cyclic consistency constraint (we describe and formulate the losses in the following subsection). Further, we show that using the above architecture, we are able to successfully perform the task of multi-domain conditional translation with improved results for the semantic scene understanding task (described in Sec. 5). The proposed method can then be used to translate clear-weather datasets such as Cityscapes and IDD to each of the three adverse-weather domains. As validated by the results in Table 2, a standard segmentation method trained on Cityscapes and then fine-tuned on a domain translated version of Cityscapes using the proposed dataset translation method, outperforms previous multi-domain translation methods on semantic segmentation performance on real-world datasets.

Adversarial loss. We use multi-scale discriminators to discriminate the images at different scales to better enforce the generated images to be visually similar to the target domain ($\mathbf{T}$) images ($\{T_{fog}, T_{rain}, T_{night}\} \in \mathbf{T}$). We denote the discriminator as D and the generator that draws a random vector $z_0 \in \mathbf{Z}$ from a random Gaussian distribution

and is conditioned by AdaIn and FADE as G. The Adversarial loss [21] formulation is as follows:

$$\min_G \max_D \mathcal{L}_{GAN}(D,G) = \mathbb{E}_{\widehat{T}_m \sim \mathbf{T}}[\log D(\widehat{T}_m)] + \mathbb{E}_{z_0 \sim p_z(z)}[\log(1 - D(G(z_0)))] \quad (3)$$

Perceptual Loss. We employ the perceptual loss [9, 35] to ensure that the translated image maintains content consistency and is perceptually similar despite the domain shift. Specifically, we use the perceptual loss modification discussed in [83]. In Eq.(4), Φ_i is the VGG-19 model pre-trained on ImageNet used to extract the feature maps from the i-th layer. Following [83], we employ interpolation to match the dimensions of the generated feature maps $\widehat{f}_{generated}$ and input content feature maps $f_{content}$ to fit the last thee layers of Φ enabling a slight improvement in the content preservation across feature levels.

$$\mathcal{L}_{Perceptual} = \tau(\sum_{i=1}^{N} \lambda_i \, \|\Phi_i(\widehat{f}_{generated}) - \Phi_i(f_{content})\| \,) \quad (4)$$

Multi-layer Patch-wise Contrastive loss. In addition to the GAN loss and perceptual loss, we use a multi-layer patch-wise contrastive loss [50] with a Generator Encoder $\mathbf{G}_{Encoder}$ (reuse the last three blocks from the generator; thus $\mathbf{G}_{Encoder}$ is a subset of $\mathbf{G}$) and a two-layer MLP ($\mathbb{H}_l$) to reduce dimensions similar to SimCLR [10]. The input content image and generated image are divided into 256 patches each. A query is a patch sampled from the target domain translated image (i.e, the generated image) and is associated with the patch at the same location in the input content image. This corresponding patch in the input is the "positive" and other non-corresponding patches are the "negatives" in the shared embedding space. Thus, the contrastive loss can be formulated as a (k+1)-way classification that associates the query vector with the positive patch and disassociates the patch with respect to the negatives in a shared embedding space. Hence, patches $p, p^+ \in \mathbb{R}^D$ and $p^- \in \mathbb{R}^{k \times D}$ where p is the query patch from the target domain, p^+ is the positive patch. There are k negative patches (p^-) in the D-dimensional shared embedding space.

$$\mathcal{L}_{CTR} = -\log \left[\frac{exp(p \cdot p^+/\tau)}{exp(p \cdot p^+/\tau) + \sum_{i=1}^{k} exp(p \cdot p_i^-/\tau)} \right] \quad (5)$$

CUT [50] shows that using internal patches (negatives from within the image) similar to SimCLR [10] setting demonstrates better translation performance as compared with using negatives from other images (external negatives) similar to a MoCo [24] setting. Consider the embedding vectors v_l obtained after the feature maps at l-th layer are passed through $\mathbf{G}^l_{Encoder}$ and $\mathbb{H}_l$, we get a stack of vectors $\{v_l\} = \{\mathbb{H}_l(\mathbf{G}^l_{Encoder}(\mathbf{G}(S_m)))\} \; \forall l \in 1, 2, 3 \cdots L$.

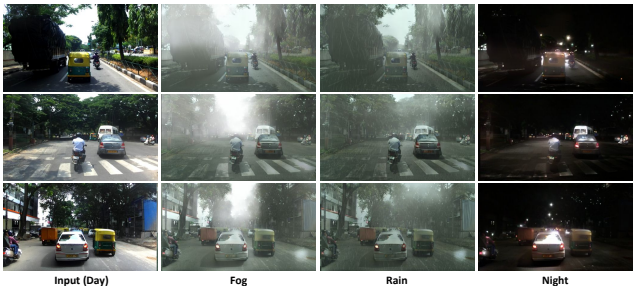


Figure 3. Qualitative results for the Indian Driving Dataset.

We denote number of patches as N_{patch}, the positive vector at layer l as v_l^+, the negatives as v_l^- and the query vector from the target domain image as $\widehat{v}_l$. Hence, we obtain the SimCLR setting patch-wise contrastive loss as follows:

$$\mathcal{L}_{InternalNCE} = \mathbb{E}_{\widehat{S}_n \sim \mathbf{S}} \sum_{l=1}^{L} \sum_{j=1}^{N_{patch}} \mathcal{L}_{CTR}(\, \widehat{v}_l, v_l^+, v_l^- \,) \quad (6)$$

Maximizing mutual information for high dimensional image spaces with a contrastive loss is equivalent to minimizing conditional entropy (InfoGAN [11] lower bound). While [50] have designed the contrastive InfoNCE loss for single domain translation, we show that using the above loss in addition to the content and style streams that capture and fuse domain-invariant and domain-specific features enables unsupervised multi-domain image-to-image translation.

Loss function. Our loss function objective for two-stream contrastive multi-domain image translation model includes the GAN loss to generate realistic images in the target domain, the modified perceptual loss to preserve source content when conditioned with style from the target domains, and the patch-wise contrastive loss to effectively disentangle style from content and enabling better content preservation in the image. The full objective is as follows:

$$\mathcal{L} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{Perceptual} + \lambda_2 \mathcal{L}_{CTR} \quad (7)$$

where λ_1 and λ_2 are hyper-parameters to weigh the losses that contribute to content preservation.

4. Experiments

In this section, we elaborate on the training and explain the evaluation pipeline. We also discuss how the Semantic Foggy Scene Understanding Task (SFSU) [20, 58] is extended to evaluate the generalization performance of a model fine-tuned on the adverse domain translated datasets.

Datasets.[1] We use a composite of images from the Cityscapes [14] and Indian Driving Dataset [70] for the source clear-weather domain. We use a composite target dataset with three domains - Fog (Foggy-Cityscapes

[1]Please refer to the supplementary material for more detail on datasets.

Figure 4. Qualitative translation results for the Cityscapes Dataset.

[15, 58]), Rain (Rainy-Cityscapes [28]) and Night-time (BDD100K Night Images [78]). For evaluation, we use the following: Fog (Foggy Zürich [16, 57], Foggy Driving [15], Foggy-Cityscapes [58]), Rain (ACDC-Rain [60], Rainy-Cityscapes [28]) and Night-time (Dark Zürich [16]).

Training. We use 9435 images for the source (composite of 2975 Cityscapes images and 6460 IDD-HQ images) and 9435 images for the target domain with 3145 images from each of the three adverse weather domains. Every epoch has 9435 steps and each step includes one image from the source composite dataset and one image from the target composite dataset that is passed to the content and style conditioning streams respectively. The Generator $\mathbf{G}$ is inverse with respect to the content and style streams and incorporates the content and style information in a progressively increasing feature map scale. The contrastive loss is computed by reusing the three blocks and feature maps of the Generator $\mathbf{G}$ as $\mathbf{G_{Encoder}}$ followed by the two-layer MLP $\mathbb{H}_i$. We use the Adam optimizer [36] to train the model for 25 epochs on an NVIDIA Tesla V100 GPU with loading size of 2048×1024 and crop size of 1024×512 taking up 18 GB of GPU VRAM. Similar to [34], we adopt a two-time update rule [26] to train $\mathbf{G}$ and the multi-scale discriminators for convergence and also introduce Spectral Norm [46] in all layers to enforce Lipschitzness and improve training stability. The spatial resolution of the generated images are same as that of the Cityscapes dataset (2048×1024).

Evaluation. We evaluate our method using the semantic scene understanding downstream task to validate the generalization capability of a pre-trained model fine-tuned on the domain-translated dataset. We follow the Semantic Foggy Scene Understanding task (SFSU) [20, 22, 58] and extend the premise to Night-time and Rainy scene understanding. An outline of the steps for evaluation is as follows:

1. First, we train a semantic segmentation model ($\mathbb{M}$) on the Cityscapes dataset (clear-weather hence we denote it as $\mathbb{D}_{clear}$). For this work, we use the recent work DDRNet [27] (DDRNet-39 with ResNet-34 backbone [25]) because the method demonstrates the best semantic segmentation performance (mIoU) in Table 1.

2. Then, we use an image translation method (see Table 2) to perform dataset translation of Cityscapes ($\mathbb{D}_{clear}$) to the three target domains to obtain $\mathbb{D}_{fog}$, $\mathbb{D}_{rain}$ and $\mathbb{D}_{night}$ as the fog, rain and night domain-translated datasets respectively.

3. We then make a copy of the $\mathbb{D}_{clear}$ trained $\mathbb{M}$ network for each of the three target domains and fine-tune them respectively on $\mathbb{D}_{fog}$, $\mathbb{D}_{rain}$ and $\mathbb{D}_{night}$ to obtain $\mathbb{M}_{fog}$, $\mathbb{M}_{rain}$ and $\mathbb{M}_{night}$ for the image translation method used in Step 2.

4. We validate the fine-tuned models on real-world datasets for semantic segmentation. We evaluate the $\mathbb{M}_{fog}$ network on Foggy Zürich [58], Foggy Driving [15] and Foggy-Cityscapes [15, 58], $\mathbb{M}_{rain}$ network on ACDC-Rain [60] and Rainy-Cityscapes [28], and the $\mathbb{M}_{night}$ network on the Dark Zürich [57] dataset.

As tabulated in Table(1), the DDRNet-39 model ($\mathbb{M}$) has an mIoU performance of 79.63 on the Cityscapes test-set, but when evaluated on Foggy/Rainy/Night datasets, there is a significant decrease in performance: mIoU of 40.85 for Fog, mIoU of 41.37 for Rain and only 14.61 for Night-time scenes. Following Steps 1-4 above, the evaluation of fine-tuned $\mathbb{M}$ networks for recent single domain and multi-domain image translation methods is tabulated in Table 2.

Baselines. We consider several classic and recent image translation models to compare with our method for the dataset translation (enrichment/augmentation) task. These models will be trained on the same datasets as our method (single domain translation methods: CycleGAN [84] and CUT [50] will have one model for each adverse-weather domain) and be used for the dataset translation task.

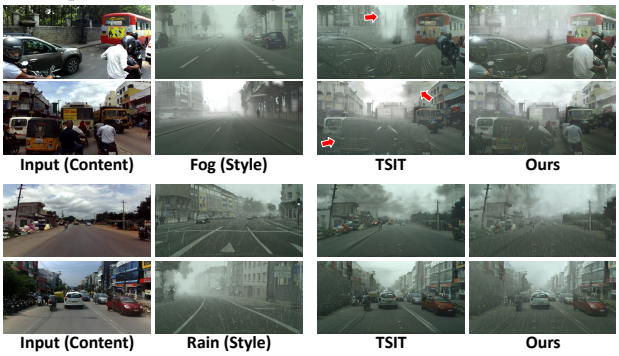
Table 2. Comparison of performance for semantic segmentation models pre-trained on Cityscapes and fine-tuned on the respective domain-shifted datasets. The fine-tuned models are evaluated on the adverse-weather datasets (in Mean IoU). Note: For the single domain image translation models - CycleGAN and CUT, we train one model for each domain (Fog, Rain, Night-time). The best results are in bold and second-best is underlined. From Table 1, DDRNet-39 trained on Cityscapes has a Mean IoU of 79.63 on the Cityscapes test-set.

Semantic Segmentation: DDRNet-39 Trained on Cityscapes and Fine-tuned on Domain-shifted Data.						
Fine-tuning (↓)	Foggy Zürich	Foggy Driving	Foggy-Cityscapes	ACDC-Rain	Rainy-Cityscapes	Dark Zürich
Cityscapes [14]	32.55	40.85	62.47	41.37	45.73	14.61
CycleGAN [84]	38.45	43.25	67.85	43.67	57.12	31.50
MUNIT [31]	37.12	43.55	67.29	42.71	56.74	29.94
DRIT++ [38]	38.63	44.18	68.45	44.05	59.28	30.36
CUT [50]	37.48	44.07	<u>68.52</u>	44.29	60.17	32.02
TSIT [34]	<u>39.81</u>	<u>44.59</u>	68.02	<u>44.41</u>	<u>62.39</u>	<u>33.46</u>
Proposed	**41.69**	**45.35**	**69.74**	**44.56**	**63.19**	**35.36**

Table 3. Semantic segmentation generalization performance (in mIoU) when translation models are trained with 501 total images (167 images each from Foggy-Cityscapes, Rainy-Cityscapes and BDD100K-Night) following the fine-tuning strategy as Table 2.

Fine-Tune (↓)	Foggy Zürich	Rainy-Cityscapes	Dark Zürich
TSIT [34]	34.11	54.89	29.64
Proposed	**34.98**	**56.12**	**30.02**

Figure 5. Disentangling Fog and Rain from Foggy-Cityscapes and Rainy-Cityscapes. Methods have been trained on 167 images from each target domain in a multi-domain setting. For fog translation, the TSIT results have an overall grey effect with loss in content information and a stronger influence of style as compared with the proposed method. For the rain translation, the TSIT results have very few rain droplets in the image whereas our result has both rain droplets (streaks) and fog at a distance (from the camera).



5. Results and Discussion

The results for the semantic scene understanding task is in Table(2). We see that our method outperforms other single-domain and multi-domain image translation methods including TSIT [34], DRIT++ [38] and MUNIT [31]. The performance improvement of the proposed method over multi-domain translation methods TSIT and DRIT++ especially for the Fog and Rain datasets is in part due to the better disentanglement of rain and fog styles when trained with a composite of Foggy-Cityscapes and Rainy-Cityscapes (and BDD100K-Night). This is because the synthetic rain images in Rainy-Cityscapes dataset has been created based on the visual effects of rain in real-world images using scene depth information to synthesize rain streaks and fog as a function of distance from the camera. Since both domains have synthetic fog at different intensities, it is thus necessary to effectively disentangle the rain and fog styles. In Fig.(4), we show a qualitative comparison of the translation results on the Cityscapes dataset highlighting the artefacts in the translation results using baseline methods (red arrows). From the image, we notice the completely dark regions above the vehicles for MUNIT and DRIT++, whereas the result from the proposed method preserves the objects from the daylight scene such as the road sign and also includes realistic street-lighting. In addition, the translated images using MUNIT, DRIT++ and TSIT for the rain domain lacks rain streaks whereas the translation result for the proposed method includes rain streaks and appropriate fog. We attribute the higher performance of the proposed method over the baselines on the semantic scene understanding task to the improved content preservation and style infusion in the proposed formulation. Further, the proposed method also outperforms single-domain translation methods: CycleGAN and CUT, despite the baselines having one translation model for each domain as compared to the proposed multi-domain translation method where only one model is trained on the composite target dataset ($\mathbf{T}$).

To investigate the performance of our method when using only as few as 167 images from each target weather domain (the source and target have a total of 501 images each), we find better style infusion and content preservation in the translated results from the proposed method as compared with TSIT. Quantitative results following the same evaluation strategy as described in Sec.(5) for the low-data regime are in Table 3. In the qualitative comparison in Fig.(5), the translation results for TSIT shows a loss in content information and is more strongly influenced by the style. As de-

Figure 6. Qualitative results of ablation experiments.

Table 4. Quantitative evaluation of ablation experiments measuring Mean FID (lower is better) for the multi-domain image translation task (fog, rain, night) on the Cityscapes validation-set.

Sl.	Ablation Experiment	Mean FID ↓
1.	w/o Target style conditioning	82.352
2.	w/o AdaIn	82.934
3.	w/o Denormalization	81.799
4.	w/o L_{CTR}	85.943
5.	w/o $L_{Perceptual}$	157.401
6.	Proposed	**79.806**

scribed in Sec.(3.2) we use a modified perceptual loss and a multi-layer patch-wise contrastive loss to better preserve features from the content image and infuse style progressively, hence we observe better translation results with the proposed method as compared with TSIT with lesser data.

We show qualitative results[2] using the proposed method for the Indian Driving Dataset (IDD) [70] in Fig.(4). As discussed in Sec.(2), the IDD dataset shows a scenic domain-shift compared to the European driving scenes in Cityscapes, yet the results using the proposed method show realistic adverse-weather degradation.

Ablation experiments.[3] For multi-domain image translation methods, the network design and implementation choices are important for effective training. We conduct ablations on the different components of the proposed method to investigate the diversity of translation. We report the results for the ablation experiments for the multi-domain translation tasks on the Cityscapes validation set using the Fréchet Inception Distance (lower is better) metric. We only report the FID metric for ablations in Table 4 and omit the Inception Score (IS) as Barratt *et al*. [3] have shown that Inception Score is not a useful evaluation metric for comparing models. From the results in Table 4, we see the lowest FID metrics indicating the higher image generation quality using the proposed approach. In Fig.(5), we show a qualitative comparison among the different ablation experiments in Table 4. We find that removing the perceptual loss leads to severe loss in source content and removing the contrastive loss leads to loss in both content structure and target domain style. The same observation is reflected in the mean MIoU for in Table 4. Further, the results (1.) without target style conditioning and (2.) without AdaIn, the method is unable to infuse any realistic style to the content and degrades the image. Similarly, providing the input as is without the content network stream leads to improper stylization.

[2]Please refer to the supplementary material for more results.

[3]Please refer to the supplementary material for a high-resolution image.

6. Conclusion

In this work, we propose a two-stream multi-domain image translation method to effectively translate a driving dataset from clear-weather to adverse-weather domains: fog, rain and night-time. Our method is able to effectively capture domain-invariant content from the content stream using an adaptive denormalization method and is also able to fuse style captured by the style stream with adaptive instance normalization at multiple feature scales. The proposed method uses the modified perceptual loss and a multi-layer patch-wise contrastive loss to disentangle and preserve content and structure from the source domain, and also utilizes multi-scale discriminators to learn the translational mapping from one domain to many domains when conditioned with the target domain image. We also show that using only 167 images for each target domain, we are able to successfully learn the high-dimensional mapping to obtain realistic translation results. With the proposed method in this work, we show improved semantic scene understanding performance across the three domains performing better than the baseline methods on real-world datasets. This validating the hypothesis that the proposed multi-domain translation method can be used to translate clear-weather datasets to adverse-weather domains to help improve the performance of vision systems for autonomous navigation in unstructured and challenging real-world scenarios. We hope that our work will benefit the data augmentation and dataset enrichment community and spur new research directions to augment and improve existing datasets in addition to spending resources in compiling new data.

Acknowledgements: This work was supported by a project grant from WIRIN (Wipro IISc Research Innovation Network). We would also like to thank the anonymous reviewers for their valuable suggestions.

References

[1] Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2018. 2

[2] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. *arXiv preprint arXiv:2006.06500*, 2020. 3

[3] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 8

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 3

[5] Hsin-Yu Chang, Zhixiang Wang, and Yung-Yu Chuang. Domain-specific mappings for generative adversarial style transfer. In *European Conference on Computer Vision*, pages 573–589. Springer, 2020. 2

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 1

[7] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017. 3

[8] 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)*, pages 801–818, 2018. 2

[9] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 5

[11] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016. 5

[12] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019. 3

[13] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[14] 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. 1, 2, 5, 7

[15] 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, 2020. 1, 3, 6

[16] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018. 1, 6

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 3

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 3

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[20] Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, and Luc Van Gool. Analogical image translation for fog generation. *arXiv preprint arXiv:2006.15618*, 2020. 5, 6

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 5

[22] Martin Hahner, Dengxin Dai, Christos Sakaridis, Jan-Nico Zaech, and Luc Van Gool. Semantic understanding of foggy scenes with purely synthetic data. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3675–3681. IEEE, 2019. 6

[23] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10203–10212, 2019. 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3, 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a

two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[27] Yuanduo Hong, Huihui Pan, Weichao Sun, Yisong Jia, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv preprint arXiv:2101.06085*, 2021. 2, 6

[28] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2019. 1, 3, 6

[29] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–731, 2018. 3

[30] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 3, 4

[31] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2, 3, 7

[32] 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. 4

[33] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8346–8355, 2020. 1, 3

[34] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *European Conference on Computer Vision*, pages 206–222. Springer, 2020. 2, 3, 4, 6, 7

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[37] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2, 3

[38] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020. 2, 7

[39] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6758–6767, 2019. 3

[40] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *arXiv preprint arXiv:1809.01361*, 2018. 3

[41] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019. 2

[42] 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. 3

[43] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5952–5961, 2019. 3

[44] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 2

[45] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[46] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 6

[47] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *arXiv preprint arXiv:1711.02536*, 2017. 3

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3

[49] Kwanyong Park, Sanghyun Woo, Inkyu Shin, and In-So Kweon. Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. In *34th Conference on Neural Information Processing Systems, NeurIPS 2020*. Conference on Neural Information Processing Systems, 2020. 3

[50] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2, 5, 6, 7

[51] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2, 4

[52] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11908–11915, 2020. 1

[53] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018. 3

[54] Doyen Sahoo, Hung Le, Chenghao Liu, and Steven CH Hoi. Meta-learning with domain adaptation for few-shot learning under domain shift. 2018. 3

[55] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *arXiv preprint arXiv:2002.07953*, 2020. 3

[56] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. *arXiv preprint arXiv:2007.07431*, 2, 2020. 3

[57] 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. 1, 3, 6

[58] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018. 1, 2, 3, 5, 6

[59] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *arXiv preprint arXiv:2005.14553*, 2020. 3

[60] 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. 1, 6

[61] Mark Schutera, Mostafa Hussein, Jochen Abhau, Ralf Mikut, and Markus Reischl. Night-to-day: Online image-to-image translation for object detection within autonomous driving by night. *IEEE Transactions on Intelligent Vehicles*, 2020. 1

[62] Vishwanath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision*, pages 763–780. Springer, 2020. 1

[63] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 3

[64] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 3

[65] Robby T Tan. Visibility in bad weather from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[66] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 3

[67] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul de Charette, and Jean-François Lalonde. Rain rendering for evaluating and improving robustness to bad weather. *International Journal of Computer Vision*, 129(2):341–360, 2021. 3

[68] 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. 3

[69] 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. 3

[70] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751. IEEE, 2019. 1, 2, 5, 8

[71] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12270–12279, 2019. 1, 3

[72] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[73] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. *arXiv preprint arXiv:2104.09367*, 2021. 1

[74] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. In *European Conference on Computer Vision*, pages 327–342. Springer, 2020. 3

[75] Wenhan Yang, Robby T Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *IEEE Transactions on pattern analysis and machine intelligence*, 2020. 3

[76] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 3

[77] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral net-

work with guided aggregation for real-time semantic segmentation. *arXiv preprint arXiv:2004.02147*, 2020. 1, 2

[78] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 6

[79] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. *arXiv preprint arXiv:1909.07877*, 2019. 3

[80] Ning Zhang, Lin Zhang, and Zaixi Cheng. Towards simulating foggy and hazy images and evaluating their authenticity. In *International Conference on Neural Information Processing*, pages 405–415. Springer, 2017. 1

[81] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5943–5951, 2019. 3

[82] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[83] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In *The IEEE European Conference on Computer Vision (ECCV)*, August 2020. 1, 2, 5

[84] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3, 6, 7