Dual-Mode Training with Style Control and Quality Enhancement for Road Image Domain Adaptation

Moritz Venator[1,2], Fengyi Shen[3], Selcuk Aklanoglu[1], Erich Bruns[1], Klaus Diepold[3], Andreas Maier[2]

[1]AUDI AG, [2]Friedrich-Alexander University Erlangen-Nürnberg, [3]Technical University of Munich

[1]{first.last}@audi.de, [2]{first.last}@fau.de, [3]{fengyi.shen, kldi}@tum.de

Abstract

Dealing properly with different viewing conditions remains a key challenge for computer vision in autonomous driving. Domain adaptation has opened new possibilities for data augmentation, translating arbitrary road scene images into different environmental conditions. Although multimodal concepts have demonstrated the capability to separate content and style, we find that existing methods fail to reproduce scenes in the exact appearance given by a reference image. In this paper, we address the aforementioned problem by introducing a style alignment loss between output and reference image. We integrate this concept into a multimodal unsupervised image-to-image translation model with a novel dual-mode training process and additional adversarial losses. Focusing on road scene images, we evaluate our model in various aspects including visual quality and feature matching. Our experiments reveal that we are able to significantly improve both style alignment and image quality in different viewing conditions. Adapting concepts from neural style transfer, our new training approach allows to control the output of multimodal domain adaptation, making it possible to generate arbitrary scenes and viewing conditions for data augmentation.

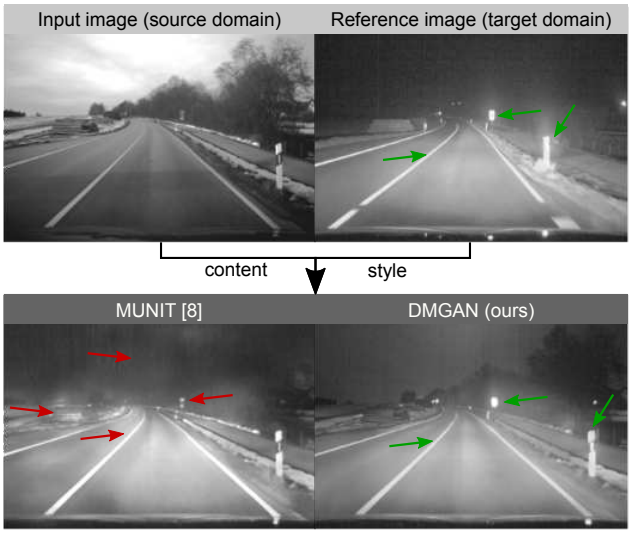


Figure 1. We introduce the concept of domain adaptation with style alignment. The output image should combine the content (viewpoint, scene structure) of the input image with the style of the reference image in the target domain. Compared to other state-of-the-art methods, our proposed method enhances image quality and style control, adopting the appearance of objects given by the reference (green arrows) and reducing artifacts (red arrows).

1. Introduction

Automated driving functions demand a reliable and precise representation of the car's environment, which is extracted by various kinds of sensors. Camera systems are employed to detect a wide range of objects being indispensable for the driving task, e.g., vehicles, pedestrians, lane markings, and traffic signs. For the development of self-driving cars, the requirements regarding performance of image processing algorithms rise tremendously [32].

Handling images captured under varying and occasionally extreme weather and illumination conditions is one of the main challenges. This applies not only to algorithms, but also to training and test data which is rare for those cases and can only be acquired with enormous effort.

Thus, the possibility of synthesizing authentic data for those rare conditions is desirable. Simulation engines are a common way to generate synthetic images [12], but often fail to reproduce real-looking images for all scenarios due to the complex interaction of scene content, i.e., the structure and semantics of the road scene, image sensor properties (e.g., the behavior in changing viewing conditions), and environmental conditions such as weather and daytime.

Recently, Generative Adversarial Networks (GANs) [5] have opened up a new research field for data augmentation, enabling image-to-image translation from one domain to another [36, 19, 10], e.g., from sunny to rainy viewing conditions. Multimodal architectures [8, 37, 16] learn dis-

entangled representations of the content, which contains the geometric and semantic structure of the image scene, and style, which encodes illumination and texture. This makes it possible to either alter the appearance of a fixed content by generating diverse images with different style codes or applying a fixed style to different contents resulting in a style-aligned set of images. In our application, this concept would allow to transfer the appearance of a reference image in order to reproduce challenging viewing conditions for arbitrary road scene contents.

However, as we show in our paper, aligning road images captured at the same location, but under different viewing conditions, in one domain and style often results in inconsistent scene appearance with a state-of-the-art image-to-image translation framework [8]. This is a result of implicit learning of content and style representations which does not guarantee a meaningful disentanglement [13].

Addressing the aforementioned problem, we extend the concept of cycle reconstruction [36] and disentangled content and style representation [8] by introducing Dual-Mode GAN (DMGAN), a modified network architecture and training setup which does not only allow multimodal and cross-domain image-to-image translation, but also explicitly enforces style alignment and improves translation quality (see Fig. 1).

Our contributions are summarized as follows:

- Proposing a style alignment loss between output and reference image, we integrate the constraint into an image-to-image translation model through a novel dual-mode training process which alternately switches between random and guided optimization steps.
- We extend the multimodal image domain adaptation model with additional adversarial losses that enhance image quality and style control.
- In our experiments, we show that our approach leads to significant improvements in style alignment, but also image quality and feature recovery in road images.

2. Related Work

Generative Adversarial Networks Recently, GANs [5] have shown impressive results in generative tasks such as image generation [26, 33], text-to-image synthesis [35, 27], video frame prediction [15, 18], and face reenactment [34]. GANs usually consist of a generator and a discriminator network. The generator aims to synthesize images that closely model the distribution of input data while the discriminator learns to distinguish real and fake images. Aiming to minimize their training losses, both networks ideally improve in their respective tasks [5].

Image-to-image translation Beside the aforementioned applications, GANs can be used for cross-domain image-to-image translation [36, 19], which describes the task of translating an image from one domain into another one. A domain may represent a collection of images captured under certain viewing conditions or with a specific sensor, e.g., photographs taken at different daytimes [19] or thermal images [10], but also other types of inputs and representations, e.g., simulated images or semantic labels [36].

The image-to-image translation can be trained with paired images, employing the concept of conditional GANs [24, 10]. Since acquisition of such paired data is often unfeasible due to scene dynamics, Zhu *et al*. proposed CycleGAN [36] for unsupervised learning. By introducing a cycle consistency constraint which compares input and cycle-reconstructed images pixel-wisely, the model can be trained with unpaired image collections. However, the image-to-image translation is deterministic, neglecting the multimodality of real-world scenes in which viewing conditions inside a domain can vary heavily [37, 16]. Therefore, the authors of MUNIT [8] extend the concept by splitting images into a content feature map and a manipulable style vector. Combined with cycle reconstruction, this disentangled representation allows synthesis of multimodal target domain images from a single input image. We extend their concept by explicitly accounting for style control and enhancing image quality with additional adversarial losses.

Domain adaptation for feature-based localization Image-to-image translation can also be employed for the task of feature-based localization [29]. Porav *et al*. [25] propose a new feature descriptor loss in the cycle reconstruction step for recovery of features lost due to changes in viewing conditions. However, the supervised fine-tuning stage introduces additional effort to collect paired data, making it unapplicable in many real cases. In the architecture of ToDayGAN [1], a discriminator architecture operating separately on blurred-RGB, grayscale, and gradient images is introduced as a solution for improving translations and place recognition with night images.

Those results show that feature matching can be used as an evaluation measure for domain adaptation performance, which we will also make use of to assess our models.

Neural style transfer Neural style transfer [4, 7, 17] has become popular in various tasks such as photo editing and fashion design. Typically, the style of a reference image, e.g., a painting, is extracted and applied to a normal image like a photograph. Convolutional Neural Networks (CNNs) such as VGG [30] have very rich internal representations of what content and style look like. Gatys *et al*. [4] have shown that those representations are independent from each other, allowing to extract style and content from the reference and

input images, respectively. Thus, applying the style properties of the reference image to the content of the input image results in the desired output.

Other than comparing gram matrices of VGG feature maps as style loss in an iterative, time-consuming optimization process [4], adaptive instance normalization (AdaIN) [7] presents a different approach to allow real-time and arbitrary style transfer. By rescaling the normalized feature map of the input image with the feature mean and variance of the style image, the network learns to alter image style and texture during training.

So far, the concept of style transfer has been mainly applied to modify style and appearance of photographs. Kazemi [13] have presented first ideas to combine style losses with adversarial losses to further improve the disentanglement of content and style for image generation. However, to the best knowledge of the authors, our work is the first approach combining cycle and style consistency for bidirectional, cross-domain image-to-image translation.

3. Dual-Mode GAN for Style Control and Quality Enhancement (DMGAN)

In this work, we consider the problem of style alignment for domain adaptation of road scene images. Given a reference style extracted from an image in the target domain, the output image should show the scene content of the input image, but resemble the style (weather, illumination, road texture, shadows, reflections, etc.) of the reference image. This would allow to translate images from arbitrary domains into a target domain and style while rendering any content given in the input image in specific viewing conditions.

The basic architecture of our model is based on MUNIT [8], a state-of-the-art method for multimodal unsupervised image-to-image translation. In their method, Huang *et al*. disentangle content and style of images from two domains while the content is shared by the domains in a mutual latent space. As a result, MUNIT is able to generate diverse image-to-image translations from one to another domain while obtaining a high visual quality.

Given an input image x_A from one domain $\mathcal{X}_A$, content c_A and style s_A are extracted by the content and style encoders, E^c_A and E^s_A. Combining the content c_A with a style vector s_B of the other domain $\mathcal{X}_B$ via a multi-layer perceptron (MLP) and AdaIN [7], the decoder G_B outputs an image $x_{A\to B}$ in domain $\mathcal{X}_B$:

$$x_{A\to B} = G_B(c_A, s_B) = G_B(E^c_A(x_A), s_B). \qquad (1)$$

By changing values of the style vector s_B, the appearance of the output image can be changed. It can be either an arbitrary vector sampled from a prior distribution $p(s_B) \sim \mathcal{N}(0, 1)$ or a reference style vector extracted from a target domain image x_B by the style encoder E^s_B.

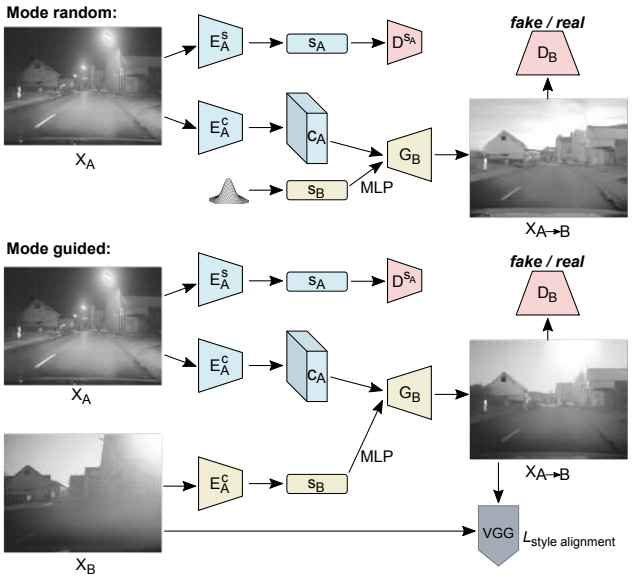


Figure 2. The training procedure is composed of two modes. For simplification, only one forward translation for each mode is shown, the opposite translation and cycle-reconstruction is similar to MUNIT [8]. During training, we switch the mode after each iteration. The images are sampled independently from each other.

In their work, Huang *et al*. [8] show that, by using a style extracted from a reference image, the respective style can be applied to the output image. As we will reveal in the results section, this does not always lead to consistent results when applied to weather and illumination conditions in road images. Our observation can be explained by the fact that MUNIT does not explicitly account for style alignment, but learns the representation in the style space implicitly by combining random sampling from a prior distribution with an adversarial network, a concept known from adversarial autoencoders (AAE) [23].

We address this problem by proposing a novel dual-mode training with alternating optimization steps employing a style alignment constraint, which is further explained in the following section. Afterwards, we explain how we adapt the composition of discriminators to further boost visual quality with styles taken from a reference image or a prior distribution.

3.1. Dual-mode training for style alignment

Unlike MUNIT [8], which exclusively samples random style vectors from a prior distribution during training, we propose an alternating, dual-mode approach, switching the sampling of the style vectors after each training iteration as visualized in Fig. 2. In the first mode, which we call *random mode*, we keep the generator concept of MUNIT [8] that allows to sample and interpolate various styles following a prior distribution. In the second mode, the *guided mode*, the style encoder E^s_B extracts a style vector s_B from a target

domain image x_B which shall guide the translation. This step explicitly trains the network to adopt reference styles given by other images instead of using random styles only.

Moreover, with the introduction of the guided mode, we are able to introduce a *style alignment loss* $\mathcal{L}_{\text{sa}}$ enforcing style similarity between the translation output and the reference image while we still stay in an unsupervised training setup. Modifying a concept from style transfer with AdaIN [7], a subset of layers from a pretrained VGG16 [30] network is taken for the style alignment loss (see Fig. 3), which is computed by comparing mean μ and variance σ of the output feature maps ϕ between two images. Our experiments have revealed that putting equal weight to all layers decreases the visual quality. Instead, we increase the weights for deeper layers to emphasize semantic, object-level features [22]. The loss is summed up for all layers:

$$\mathcal{L}_{\text{sa}}^{x_{A\to B}} = \sum_{i=1}^{L} w_i(\|\mu(\phi_i(x_{A\to B})) - \mu(\phi_i(x_B))\|_1 + \|\sigma(\phi_i(x_{A\to B})) - \sigma(\phi_i(x_B))\|_1), \quad (2)$$

where ϕ_i is the corresponding feature map of the ith selected layer, $x_{A\to B}$ represents the translated output, and x_B the reference style image. The weight for each selected VGG layer is empirically set to $w_i = 0.2 \cdot 2^{i-1}$.

Similar to MUNIT [8], our model is trained with four types of reconstruction losses: The image reconstruction loss (3) trains the convolutional autoencoder for image recovery, the content (4) and style (5) reconstruction losses enforce the reconstruction of the latent code when encoding the translated image $x_{A\to B}$ again:

$$\mathcal{L}_{\text{recon}}^{x_A} = \|G_A(E_A^c(x_A), E_A^s(x_A)) - x_A\|_1, \quad (3)$$
$$\mathcal{L}_{\text{recon}}^{c_A} = \|E_B^c(G_B(c_A, s_B)) - c_A\|_1, \quad (4)$$
$$\mathcal{L}_{\text{recon}}^{s_B} = \|E_B^s(G_B(c_A, s_B)) - s_B\|_1. \quad (5)$$

The cycle reconstruction loss ensures that the backward translation $x_{A\to B\to A}$ is consistent with the input image x_A:

$$\mathcal{L}_{\text{cyc}}^{x_A} = \|G_A(E_B^c(G_B(E_A^c(x_A), s_B)), E_A^s(x_A)) - x_A\|_1. \quad (6)$$

Additionally, a domain-invariant perceptual loss [11] $\mathcal{L}_{\text{percep}}^{x_{A\to B}}$ is employed to allow similar high-level convolutional features to be detected in both output $x_{A\to B}$ and input image x_A. Since the perceptual loss is already used in MUNIT and also based on VGG [30], we can reuse the output of the forward pass so that there is no slowdown in training due to the newly introduced style alignment loss.

3.2. Discriminator architecture

The original discriminator used in MUNIT [8] consists of three identical networks operating on different image resolutions where each judges if the input image is real or fake:

$$D(x) = D_{\text{p1}}(x) + D_{\text{p2}}(f_{\text{avg,p2}} * x) + D_{\text{p4}}(f_{\text{avg,p4}} * x), \quad (7)$$

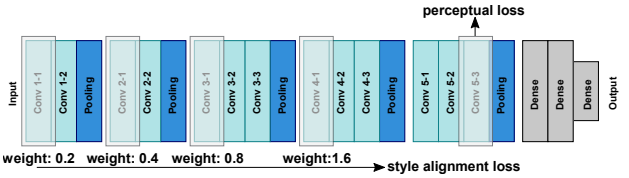
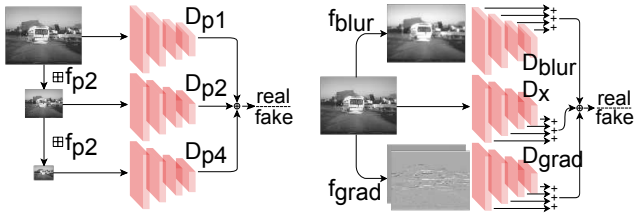


Figure 3. For computation of the style alignment loss, the differences of mean and variance from selected VGG16 [30] feature maps are compared, weighted, and summed up.



(a) MUNIT discriminator [8]. (b) ToDayGAN discriminator [1].

Figure 4. Comparison of discriminator architectures operating on different image transformations. We experiment with both variants and choose (b) in our final model.

with $f_{\text{avg,ps}}$ being an average pooling filter with stride s.

In our model, we replace the multi-scale architecture of the discriminator by an approach introduced in ToDayGAN [1], which transforms the output image into three different representations and sends them to separate copies of the same sub-network. With an output after each convolutional layer, each network assesses the input at multiple patch levels (see Fig. 4) with ascending weight.

The first transformation is a blurred version of the translated output, which helps to distinguish brightness, contrast, and major illumination differences according to [9]. The second one – in our application – is the identity transformation, allowing the discriminator to give a prediction based on image texture. The third transformation is the concatenation of horizontal and vertical gradients, assisting the discriminator to differentiate the appearance of edges.

The final decision is obtained by summing up the outputs of the three networks:

$$D(x) = D_x(x) + D_{\text{blur}}(f_{\text{blur}} * (x)) + D_{\text{grad}}(f_{\text{grad}} * x). \quad (8)$$

The different input types allow the networks to focus on multiple aspects, making it easier to discriminate between real and fake images in the respective domains.

So far, our adversarial loss for the generator and discriminator can be formulated as

$$\mathcal{L}_{\text{adv}}^{G_B} = (D_B(G_B(c_A, s_B)) - 1)^2, \quad (9)$$
$$\mathcal{L}_{\text{adv}}^{D_B} = (D_B(G_B(c_A, s_B)))^2 + (D_B(x_B) - 1)^2. \quad (10)$$

Moreover, during training, we introduce a style discriminator D_A^s following the AAE concept [23] to judge whether

an encoded style vector s_A is close to the prior distribution $\mathcal{N}(0, I)$ the styles are sampled from in the random mode. The discriminator is comprised of a simple multi-layer perceptron with the following adversarial losses:

$$\mathcal{L}_{\text{adv}}^{E_A^s} = (D_A^s(E_A^s(x_A)) - 1)^2, \tag{11}$$

$$\mathcal{L}_{\text{adv}}^{D_A^s} = (D_A^s(E_A^s(x_A)))^2 + (D_B^s(s'_A) - 1)^2, \tag{12}$$

where s'_A is sampled from the prior distribution $p(s_A)$. Through the additional discriminator, we force our model to encode styles from real images in the same latent space as $p(s_A)$, expecting a more consistent mapping.

3.3. Training losses

In sum, the total generator loss for the guided mode is:

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{G,\text{total}} &= \lambda_x \mathcal{L}_{\text{recon}}^{x_A} + \lambda_c \mathcal{L}_{\text{recon}}^{c_A} + \lambda_s \mathcal{L}_{\text{recon}}^{s_B} \\ &+ \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}}^{x_A} + \lambda_{\text{percep}} \mathcal{L}_{\text{percep}}^{x_{A\to B}} + \lambda_{\text{sa}} \mathcal{L}_{\text{sa}}^{x_{A\to B}} \\ &+ \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^{G_B} + \lambda_{s,\text{adv}} \mathcal{L}_{\text{adv}}^{E_A^s}.\end{aligned} \tag{13}$$

For the random mode, the style alignment loss $\mathcal{L}_{\text{sa}}$ is omitted. For both modes, the total discriminator loss is:

$$\mathcal{L}_{\text{adv}}^{D,\text{total}} = \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^{D_B} + \lambda_{s,\text{adv}} \mathcal{L}_{\text{adv}}^{D_A^s}. \tag{14}$$

The loss functions for the opposite translation from $\mathcal{X}_B$ to $\mathcal{X}_A$ can be obtained by adapting the above equations.

4. Experimental Setup

4.1. Datasets

The main dataset we use for our experiments consists of grayscale images which we recorded with an automotive monocular front camera system with 1280 x 960 px resolution and a 46° horizontal field of view. The images were captured on 26 repetitive test drives on a 43 km route on public roads including highways, rural roads, and urban areas in changing environmental conditions (daytime, weather etc.) which we assign to different categories.

For evaluation of our proposed method, we select three domains, sun, rain, and night (see Table 1), and train our model with two condition pairs: **sun–rain** and **sun–night**. For validation, we create a test set by defining ten test spots along the track with a radius of 150 m each and removing all images within those spots from training data. Therefore, at each spot, our test set contains images of multiple domains, which can later be used to assess translated image quality by taking image pairs at similar viewpoints for experiments with style alignment and feature matching.

Another reason for choosing our own dataset is that we can also evaluate the performance on images captured in non-urban scenes such as rural roads and highways, which are often neglected in public datasets [29] although changing viewing conditions often have a higher impact on their appearance. Moreover, the image sensor used is very robust to changing viewing conditions such as weather or illumination, requiring the network to focus on more details.

Table 1. For training and evaluation, we select images from three domains in our dataset: sun, rain, and night.

Domain name	Viewing conditions	Test drives	Training images	Test images
sun	sunny daytime	6	126799	24689
rain	rainy daytime	2	46672	7352
night	dry nighttime	5	42512	19414

In addition, we also conduct experiments for the same domain pairs with sequences taken from the Oxford RobotCar Dataset [21] to assess the impact of our improvements on an urban dataset captured with a more illumination-sensitive RGB image sensor.

4.2. Training

Following the approach described in Section 3, our models are trained by switching between the two modes *random* and *guided* at each iteration. Adam optimizer [14] is adopted with $\beta_1 = 0.5, \beta_2 = 0.999$, and the initial learning rate is set to 0.0001. For all our models, batch size is set to one and weights of the losses are chosen as follows: $\lambda_{\text{adv}} = 1$, $\lambda_{\text{s,adv}} = 5$, $\lambda_x = \lambda_{\text{cyc}} = 10$, $\lambda_c = 1$, $\lambda_s = 2$, $\lambda_{\text{sa}} = 0.5$, and $\lambda_{\text{percep}} = 3$.

In all experiments, we resize the image height to 256 px while fixing the aspect ratio. During training, random crop of 256 x 256 px is applied to each sample. The output size for inference is 340 x 256 px. We trained our networks on a Nvidia Titan X GPU for at least 400,000 iterations, and a well-trained model takes about a week.

4.3. Evaluation Methods and Baselines

Style alignment As explained in Section 3, style alignment requires precise extraction of a reference style from a given image and its application to an input content. To allow convenient comparison and evaluation, we choose reference images of the target domain which have been captured at a nearby viewpoint.

Since style alignment is difficult to measure, we compare the performance between the state-of-the-art model MUNIT [8] and our proposed model with a survey which is conducted among experts who possess at least three years of professional experience in the field of image processing. The participants are confronted with 80 pairs of translated images (20 per translation direction) and are asked to choose the output which better renders the content of the source image in the style of the reference image (arrangement similar to Fig. 5). The answer results in 'no preference' if a candidate fails to decide within ten seconds.

GAN image quality metrics GAN research has put a lot of effort into the topic of measuring similarity between generated and real data distributions. Similar to other work on image synthesis, we calculate the following evaluation metrics for image quality assessment:

- **Inception Score (IS)** is a widely adopted metric [28] based on the Inception network [31] to measure realism for a generated set of images by considering two criteria, image quality (indicated by low class entropy for an individual image) and diversity (high entropy for the overall distribution).
- **Mode Score (MS)** [2] is an improved version of Inception Score which evaluates both visual quality and variety of generated images in one metric, meanwhile being able to detect mode collapse.
- **Fréchet Inception Distance Score (FID)** [6] compares the statistics of synthesized and real images, instead of measuring generated samples solely, and has shown to be well consistent with human judgement of visual quality. A lower FID corresponds to a more similar distribution between real and generated data.

To better correlate with style alignment and to verify whether our trained model is capable of generating high-quality images when taking real image styles, we extract styles of reference images from all ten spots and apply them to source domain images for translation. For each domain pair at each of the ten spots, we pick 100 images per domain for inference, therefore, each evaluation score is obtained over 1000 images in total.

Feature matching Inspired by previous work [25, 1], we further evaluate the quality of domain adaptation and style alignment with feature matching experiments. An effective translation model for road scene images is expected to recover keypoints and improve feature matching when images are aligned in good viewing conditions. We experiment with different descriptors for feature detection and matching, but only report results for SIFT [20] since the results for other descriptors are similar. RANSAC [3] is used for filtering of inlier matches.

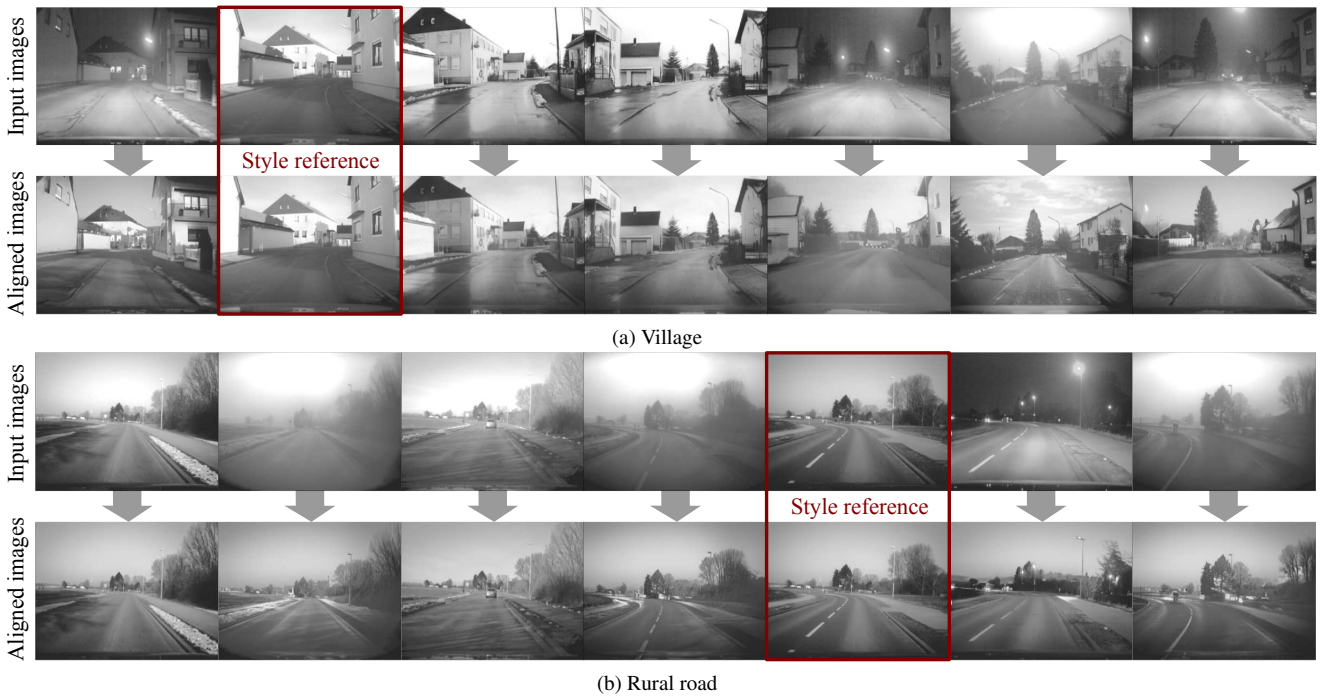
We conduct experiments on two-view matching between images captured in two different domains, but at a nearby position, comparing the results obtained with original image pairs that were not preprocessed and pairs where the image captured under bad conditions (rain/night) has been translated to sunny conditions. Beside our model, we use the unimodal UNIT [19] and multimodal MUNIT [8] architecture. For the multimodal models, we experiment with different style vectors: *zero* (mean of prior distribution), *random* (sampling from prior distribution), and *extract* (style given by the sunny reference image).

Figure 5. Style alignment comparison of MUNIT [8] and our model DMGAN using image pairs captured at similar (odd rows) and different (even rows) viewpoints. The reference image from the target domain is used to guide the appearance of the translated image. Our model improves the rendering of street and sky texture, lane markings, vegetation, and objects given by the reference image while it also reduces artifacts in the output.

5. Results

5.1. Style Alignment

For visual comparison, Fig. 5 contrasts the output of MUNIT [8] and our model. The strong similarity between output and reference image shows that our proposed model deals better with aligning image styles. Specifically, when translating from poor to good viewing conditions, our proposed approach is capable of recovering objects (e.g., road texture, lanes, vehicles, and trees) with finer edges and



(a) Village

(b) Rural road

Figure 6. Domain and style alignment of multimodal, multi-domain image collections. The upper rows depict images captured on the same street section, but at different viewing conditions. The rows below show the same images aligned in domain sun by our image-to-image translation models rain→sun and night→sun using the style extracted from a reference image (marked by red frame).

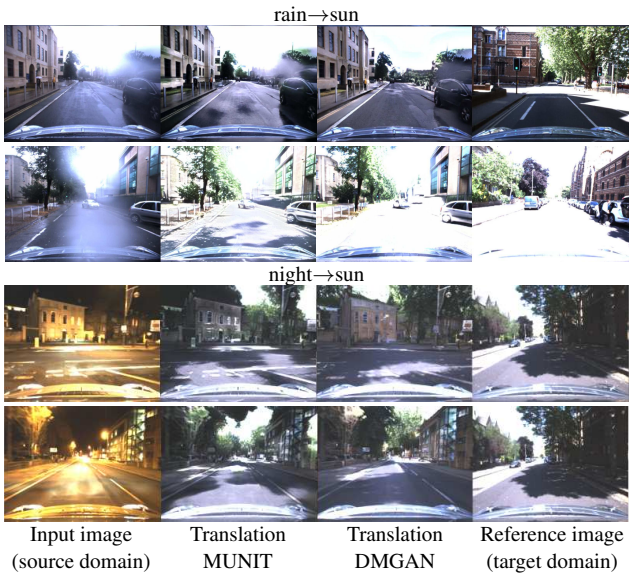


Figure 7. Domain adaptation with style alignment for images taken from the Oxford RobotCar Dataset [21]. Instead of using images captured at a similar viewpoint, the reference images are randomly selected. The outputs of MUNIT [8] and our proposed model are compared in the second and third column. Our method improves rendering of shadows, buildings, and vegetation, among others.

clearer image contents, rendering the scene in the appearance (illumination, weather, road texture, etc.) of the target image while preserving source image content as much as possible. In the survey, whose results are depicted in Table 2, our model outperforms MUNIT [8] by more than 14 % for both sun–rain translations, and by more than 54 % for the more challenging sun–night dataset. In experiments with the Oxford RobotCar Dataset, we observe similar improvements for RGB images as visualized in Fig. 7.

Table 2. Results of style alignment survey. Scores denote the percentage of generated images preferred by the human experts.

	Expert Preference		
Translation	MUNIT [8]	DMGAN	None
sun→rain	35.36 %	59.64 %	5.00 %
rain→sun	41.43 %	54.57 %	5.00 %
sun→night	18.93 %	73.57 %	7.50 %
night→sun	13.20 %	80.75 %	6.05 %

Furthermore, our model shows the capability to align images from multiple domains, thus bringing the potential to create a visually homogeneous sequence from a collection of images captured under completely different viewing conditions (see Fig. 6). The overall appearance – including illumination and weather – is well-aligned, only details such as shadows or reflections are not completely consistent.

5.2. Image Quality

Table 3 shows the GAN metrics obtained for the challenging translations from poor to sunny conditions, reveal-

Table 3. Ablation study comparing GAN metrics (explained in Section 4.3) for different model configurations. Only challenging translations from poor to good conditions are displayed here.

Method	rain→sun			night→sun		
	IS	MS	FID	IS	MS	FID
MUNIT [8]	3.492	3.277	0.190	2.707	2.793	0.274
DMGAN w/o D[a]	3.663	3.296	0.218	**3.301**	2.943	0.266
DMGAN w/o SA[b]	3.042	3.387	**0.175**	2.770	3.231	0.201
DMGAN (ours)	**3.749**	**3.405**	0.183	3.277	**3.285**	**0.198**

[a]discriminator taken from [1] [b]dual-mode training + style alignment

Table 4. Results of two-view matching before and after preprocessing with different image-to-image translation methods. Values represent average results for all ten test spots. For better comparability of multimodal models, different style types are provided.

Preprocessing Method	(Style)	rain→sun matches[a]	rain→sun inliers[b]	night→sun matches[a]	night→sun inliers[b]
None	—	316.29	37.92	286.73	22.39
UNIT [19]	—	461.55	41.26	457.78	29.36
MUNIT [8]	*zero*	470.76	42.19	431.91	28.53
	extract	488.46	42.97	465.60	29.04
	random	462.95	42.60	470.41	28.96
DMGAN (ours)	*zero*	**501.22**	**43.27**	383.49	27.40
	extract	483.40	43.21	**519.57**	**31.51**
	random	480.19	40.91	477.27	29.71

[a]SIFT feature correspondences [b]verified by RANSAC

ing that our proposed model outperforms MUNIT [8] in all quality assessment criteria. The ablation study with different model configurations shows that both the dual mode training with style alignment loss and the modified discriminator architecture contribute to the improved quality.

The example provided in Fig. 8 visualizes the effect of the two main contributions proposed in Section 3. The modified discriminator leads to sharper, clearer objects, while the style alignment enhances texture of the road surface, objects, and background. The best results with increased visibility of objects and background structures and less artifacts are obtained by the combination of both methods.

5.3. Image Feature Matching

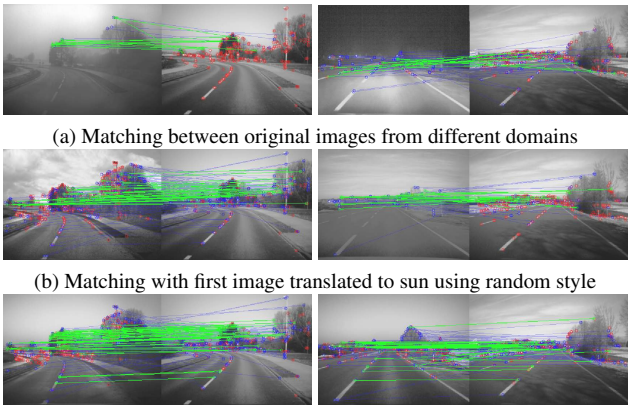
Table 4 summarizes the results of feature matching for both rain→sun and night→sun translations. By comparing results with the deterministic image translation model UNIT [19] as an additional baseline other than original images, we observe that multimodal architectures can be superior in recovering features if we control and utilize the disentangled representation. To better assess the influence of style selection on feature matching, different style types (*zero style, extracted style and random style*) are provided.

Our proposed model achieves the best matching results in both datasets. Moreover, we see a clear improvement between random and extracted styles with our method,

(a) Input image (b) DMGAN output

(c) DMGAN without modified discriminator (d) DMGAN without style alignment

Figure 8. Comparison of image quality for different model configurations. For all night→sun translations, the same style was used.



(a) Matching between original images from different domains

(b) Matching with first image translated to sun using random style

(c) Matching with first image translated to sun using extracted style

Figure 9. Visualization of cross-domain feature matching (see Table 4) comparing results before and after translation with our model. In (b) and (c), the first images (left: rain – right: night) are translated to the domain with better viewing conditions (sun). Extracting the style for the translation from the reference image (c) results in the best and most inlier matches (green lines).

demonstrating the effect of the dual-mode training for style alignment. Some examples of enhanced feature matching are visualized in Fig. 9.

6. Conclusion

In this paper, we have introduced DMGAN, a method for multimodal domain adaptation integrating a style alignment loss inspired by neural style transfer into a novel dual-mode training concept for bidirectional image-to-image translation. The improvements in style control and visual quality are verified in various experiments with road scene images including an expert survey, GAN quality metrics, and feature matching, among others. Our new approach allows to generate images of arbitrary scenes in specific viewing conditions, which is essential for data augmentation in computer vision development for autonomous driving.

References

[1] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. van Gool. Night-to-Day Image Translation for Retrieval-based Localization. In *IEEE International Conference on Robotics and Automation (ICRA)*. 2019. 2, 4, 6, 8

[2] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. Mode Regularized Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2016. 6

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 6

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1, 2

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 6

[7] X. Huang and S. J. Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510. 2017. 2, 3, 4

[8] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal Unsupervised Image-to-Image Translation. In *European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[9] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. van Gool. WESPE: Weakly Supervised Photo Enhancer for Digital Cameras. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 4

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[11] J. Johnson, A. Alahi, and F.-F. Li. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *European Conference on Computer Vision (ECCV)*, 2016. 4

[12] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan. Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks? In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 1

[13] H. Kazemi, S. M. Iranmanesh, and N. M. Nasrabadi. Style and Content Disentanglement in Generative Adversarial Networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 2, 3

[14] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2014. 5

[15] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic Adversarial Video Prediction. *arXiv:1804.01523 [cs.CV]*, 2018. 2

[16] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse Image-to-Image Translation via Disentangled Representations. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[17] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal Style Transfer via Feature Transforms. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[18] X. Liang, L. Lee, Wei Dai, and E. P. Xing. Dual Motion GAN for Future-Flow Embedded Video Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[19] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 1, 2, 6, 8

[20] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 6

[21] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 5, 7

[22] A. Mahendran and A. Vedaldi. Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016. 4

[23] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow. Adversarial Autoencoders. *International Conference on Learning Representations (ICLR)*, 2016. 3, 4

[24] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784 [cs.LG]*, 2014. 2

[25] H. Porav, W. Maddern, and P. Newman. Adversarial Training for Adverse Conditions: Robust Metric Localisation using Appearance Transfer. In *IEEE International Conference on Robotics and Automation (ICRA)*. 2018. 2, 6

[26] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs.LG]*, 2015. 2

[27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning*, 2016. 2

[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 6

[29] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018. 2, 5

[30] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs.CV]*, 2014. 2, 4

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[32] W. Wachenfeld and H. Winner. The Release of Autonomous Vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, editors, *Autonomous Driving*, pages 425–450. Springer Nature, Berlin Heidelberg, 2016. 1

[33] X. Wang and A. Gupta. Generative Image Modeling using Style and Structure Adversarial Networks. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[34] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy. ReenactGAN: Learning to Reenact Faces via Boundary Transfer. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[35] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(8), 2019. 2

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[37] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 1, 2