ManiFest: manifold deformation for few-shot image translation

Fabio Pizzati[1,2], Jean-François Lalonde[3], Raoul de Charette[1]

[1]Inria, [2]VisLab, [3]Université Laval
{fabio.pizzati, raoul.de-charette}@inria.fr, jflalonde@gel.ulaval.ca

Abstract. Most image-to-image translation methods require a large number of training images, which restricts their applicability. We instead propose ManiFest: a framework for few-shot image translation that learns a context-aware representation of a target domain from a few images only. To enforce feature consistency, our framework learns a style manifold between source and additional anchor domains (assumed to be composed of large numbers of images). The learned manifold is interpolated and deformed towards the few-shot target domain via patch-based adversarial and feature statistics alignment losses. All of these components are trained simultaneously during a single end-to-end loop. In addition to the general few-shot translation task, our approach can alternatively be conditioned on a single exemplar image to reproduce its specific style. Extensive experiments demonstrate the efficacy of ManiFest on multiple tasks, outperforming the state-of-the-art on all metrics. Our code is avaliable at https://github.com/cv-rits/ManiFest.

Keywords: Image-to-image translation, few-shot learning, generative networks, night generation, adverse weather

1 Introduction

Image-to-image translation (i2i) frameworks are gaining traction on multiple applications such as autonomous driving [24, 45] as well as photo editing [36, 37]. Those methods rely on the availability of large-scale datasets, and as such they are restricted to applications where large quantities of images are available. Unfortunately, it is unrealistic to impose significant data collection constraints every time a new i2i scenario is pursued. In addition to the complex logistics involved in acquiring large quantities of images, some scenarios may be rare (*e.g.*, auroras) or dangerous (*e.g.*, erupting volcanoes) thereby preventing even the capture of sufficient training data. Existing methods have been proposed to alleviate the requirement for large datasets, but they mostly show realistic results in highly structured environments such as face translation [29, 44, 19]. In this context, we propose ManiFest, a framework for few-shot image-to-image translation which is shown to be robust to highly unstructured transformations such as adverse weather generation or night rendering. Our approach, illustrated
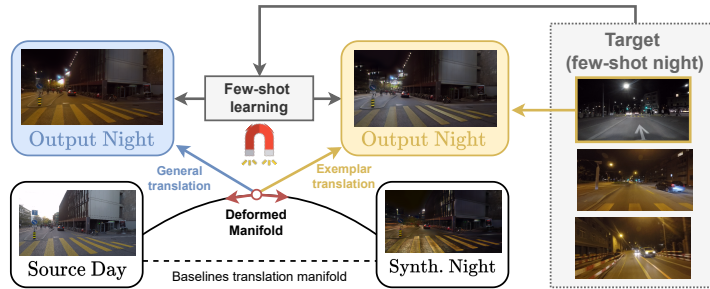
Fig. 1: Overview of ManiFest, which translates images from a source domain (here, Day) to a few-shot target (Night). Our framework learns a manifold between *anchor* domains, in the example spanning the translation between Day and Synthetic Nighttime. Our system deforms the manifold by injecting the few-shot domain information between anchor style representations, and further departs from the deformed manifold by learning to approximate the target domain *general* appearance, or to reproduce the style of a particular *exemplar*.

in Fig. 1, starts from the observation that features consistency (*i.e.* which image parts should be translated together) is crucial for unstructured i2i [31] and that the few-shot domain offers little cues to train efficiently without overfitting [34]. Indeed, rather than directly addressing few-shot i2i, ManiFest exploits features learned on a stable manifold for the few-shot domain transformation. To do so, it leverages techniques inspired from style transfer and patch-based training. We learn either to translate to some *general* style approximating the entire few-shot set, or to reproduce a specific *exemplar* from it. In short, our contributions are:

- ManiFest, a few-shot image translation framework using feature consistency by weighted manifold interpolation (WMI) and local-global few-shot loss (LGFS).
- We introduce GERM, a novel residual correction mechanism for enabling general and exemplar translation, that also boosts performances.
- Our framework outperforms previous work on adverse weather and low-light few-shot image translation tasks. We also present qualitative evaluations on rare (auroras) and dangerous (volcanoes) events.

We discuss related works in Sec. 2 and present our method in Sec. 3. The latter is thoroughly evaluated in Sec. 4 and we show several extensions in Sec. 5.

2 Related work

Image-to-image translation (i2i)   Although the early i2i translation methods required paired data [14, 64], cycle-consistency [63, 28] or recent alternatives with contrastive learning [35] have lifted such constraint. Many approaches separate style and content to enable multi-modal or multi-target translations [13, 18, 16, 5, 57], while others use additional strategies to increase scene contextual preservation [62, 15]. Translation networks can be conditioned on a

variety of additional information, including semantics [20, 42, 48, 4, 66, 65, 25, 31], instances [33], geometry [55], models [41, 40, 39, 11, 50], low-resolution inputs [1] or exemplar images [31, 59, 58]. Still, all require a large amount of data.

GANs with limited data   There have been several attempts to overcome the large data requirement for training GANs. Some use transfer learning [49] to adapt previously-trained networks to new few-shot tasks [54, 23]. In particular, [34] uses a patch-based discriminator to generalize to few-shot domains. However, these methods are designed for generative networks and do not immediately apply to i2i. Another line of work focuses on the limited data scenario [38, 2, 61, 17], but usually performs poorly when very few (10–15) images are used for training. Others exploit additional knowledge to enable few-shot or zero-shot learning, such as pose-appearance decomposition [53], image conditioning [8] or textual inputs [27]. FUNIT [29] and COCO-FUNIT [44] use few-shot style encoders to adapt the network behavior at inference time. Some use meta-learning to adapt quickly to newly seen domains [26]. Those methods show limited performance on highly unstructured scenarios. [7] exploits geometry for patch-based few-shot training, but only on limited domains with specific characteristics.

Neural style transfer   Style transfer could be seen as an instance of few-shot i2i, where the goal is combining content and style of two images [9]. This may result in distortions, which some work tried to mitigate [30]. The first examples of style transfer with arbitrary input styles are in [12, 21]. Others try to transfer styles in a photo-realistic manner by using a smoothing step [22] or with wavelet transforms [56]. These methods provide good results in some controlled scenarios, but they may fail to understand contextual mappings between source and style elements (*e.g.* sky, buildings, etc.) which we learn accurately.

3   ManiFest

The few-shot i2i task consists in learning a $\mathcal{S} \mapsto \mathcal{T}$ mapping between images of a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$ containing few training samples (*e.g.*, $|\mathcal{T}| \leq 25$). Fig. 2 presents an overview of our approach. We learn a style manifold in a standard multi-target GAN fashion (Sec. 3.1) from a set of domains which contain large amounts of training data. We call these domains *anchors*, and denote them $\mathbb{A}$. The idea of ManiFest is to simultaneously 1) learn a stable manifold using anchor domains and 2) perform few-shot training by enforcing the target style appearance to lie within the learned manifold. This allows to exploit additional knowledge, like feature consistency (*i.e.*, image parts to be translated together), learned on anchors. To this end, Weighted Manifold Interpolation (WMI, Sec. 3.2) exploits style interpolation to benefit from the learned feature consistency on anchors. We allow to further depart from the interpolated manifold with the General-Exemplar Residual Module (GERM, Sec. 3.3) which learns a residual image refining the overall appearance and thus enabling style transfer to the *general* few-shot style (approximating the entire set $\mathcal{T}$), or to a single *exemplar* in $\mathcal{T}$ as in [31]. We learn the appearance of $\mathcal{T}$ and inject it in the
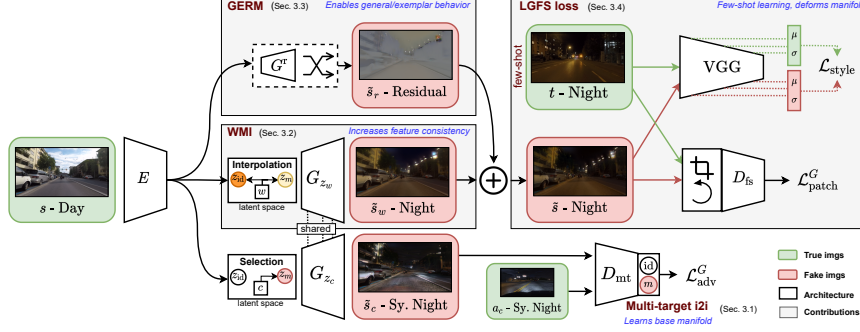
Fig. 2: ManiFest architecture, here translating Day $\mapsto$ Night using few real night images and a synthetic night anchor domain. The encoded image representation $E(s)$ is separated into content and style codes, and translated to the few-shot domain by injecting $\mathcal{T}$ on a manifold learned on anchor domains in a multi-target i2i setting (bottom). We correct the output by using residuals estimated by the GERM (top). The LGFS loss (top-right), based on statistics alignment and patch-based adversarial learning, deforms the manifold and injects $\mathcal{T}$ in it. The reconstruction cycle with a style encoder is omitted for simplicity and follows [13].

manifold with the Local-Global Few-Shot loss (LGFS, Sec. 3.4). In the following, *real* images are $s \in \mathcal{S}, t \in \mathcal{T}$, and *fake* ones $\tilde{s} \in \mathcal{T}$ where $\tilde{s}$ is our output.

3.1 Multi-target i2i

Instead of learning $\mathcal{S} \mapsto \mathcal{T}$ directly, we assume the availability of a set of two *anchor* domains, $\mathbb{A} = \{\mathcal{A}_{id}, \mathcal{A}_m\}$, with abundant data (equivalent to the "base" categories in few-shot classification, *e.g.*, [3]). By construction, one anchor is always the identity domain ($\mathcal{A}_{id} = \mathcal{S}$), while the other ($\mathcal{A}_m$) contains images easier to collect than $\mathcal{T}$, for example synthetic images or images from existing datasets. We formalize the multi-target image translation problem as learning the $\mathcal{S} \mapsto \mathbb{A}$ mapping. At training time, we disentangle image content and appearance by using content and style encoders $E(\cdot)$ and $Z(\cdot)$ respectively. We use Z for reconstruction and translation as in [13], which we refer for details. We reconstruct $s = G_{Z(s)}(E(s))$, where $G_{Z(s)}$ is the style injection of $Z(s)$ into G as in [13]. This effectively learns latent style distributions as in [13], namely here for each anchor $\{z_{id}, z_m\}$. A multi-target mechanism (following [5]) is employed in Z since we have two anchors. We translate to a randomly selected domain $c \in \{id, m\}$ with

$$z_c = [\![c = id]\!]z_{id} + [\![c = m]\!]z_m, \quad \tilde{s}_c = G_{z_c}(E(s)), \tag{1}$$

where $[\![\cdot]\!]$ are the Iverson brackets. The translation to a given anchor style is depicted in Fig. 2 as "**selection**". The multi-target discriminator D_{mt} employs adversarial losses $\mathcal{L}_{adv}^G$ and $\mathcal{L}_{adv}^D$ to force fake images $\tilde{s}_c$ to resemble $a_c \in \mathcal{A}_c$. Additional training details are in the supp. material.
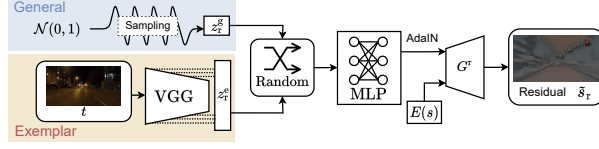
Fig. 3: GERM-based residuals. We perform either *exemplar*- or *general*-based transformations on the few-shot set by learning residuals conditioned on original image features $(E(s))$ and extracted statistics or noise, respectively. At training time, we alternate randomly the two modalities.

3.2 Weighted Manifold Interpolation (WMI)

Our intuition is that encoding $\mathcal{T}$ between the linearly interpolated style representations of $\mathbb{A}$ should enforce feature consistency in $\mathcal{T}$. For instance, assuming $\mathcal{S} = day$, $\mathcal{T} = night$, $\mathcal{A}_m = synthetic\ night$, the network will be provided with the information that all sky pixels should be darkened together.

In practice, we learn weights $w = \{w_{\mathrm{id}}, w_m\}$ which sum to 1 and encode an image $\tilde{s}_w$ with feature consistency by interpolating the anchors style representations:

$$z_w = w_{\mathrm{id}} z_{\mathrm{id}} + w_m z_m \,, \quad \tilde{s}_w = G_{z_w}(E(s)) \,. \tag{2}$$

This is visualized in Fig. 2 as **"interpolation"**. Learning w allows us to determine the point in the $\mathbb{A}$ manifold which is most consistent with $\mathcal{T}$. This point is learned with the LGFS loss (Sec. 3.4).

3.3 General-Exemplar Residual Module (GERM)

Our GERM seeks to further increase realism by learning a residual in image space. Moreover, our design enables distinguishing between *general* and *exemplar* translations. The idea is to allow deviations from the $\mathbb{A}$ manifold by learning a residual image $\tilde{s}_{\mathrm{r}}$ which helps encode missing characteristics from $\mathcal{T}$. This is done by processing the input image features $E(s)$ with a generator G^{r} such that

$$\tilde{s}_{\mathrm{r}} = G^{\mathrm{r}}_{z_{\mathrm{r}}}(E(s)) \,, \quad \text{and} \quad \tilde{s} = \tilde{s}_w + \tilde{s}_{\mathrm{r}} \,, \tag{3}$$

where z_{r} is a vector controlling general- or exemplar-based modalities. In both cases, we draw inspiration from AdaIN style injection [13] and condition the injected parameters on different vectors, as illustrated in Fig. 3.

For the *exemplar* residual, the style of a specific image $t \in \mathcal{T}$ as in [31] is reproduced by conditioning the residual on t. In this case,

$$z_{\mathrm{r}}^{\mathrm{e}} = (\mu_k(t), \sigma_k(t))|_{k=1}^{K} \,, \quad \tilde{s}_{\mathrm{r}} = G^{\mathrm{r}}_{z_{\mathrm{r}}^{\mathrm{e}}}(E(s)) \,, \tag{4}$$

where $\mu_k(\cdot) = \mu(\phi_k(\cdot))$ and $\sigma_k(\cdot) = \sigma(\phi_k(\cdot))$ are mean and variance of the k-th out of K layer outputs ϕ_k of a pretrained VGG network [12], and $|$ is the concatenation operator. Since the LGFS loss exploits VGG statistics (Sec. 3.4),

G^{r} will be driven to exploit the additional information provided by the input statistics vector, effectively making the generated image more similar to t.

We learn a *general* residual by removing the conditioning on t and by injecting random noise instead:

$$z_{\mathrm{r}}^{\mathrm{g}} \sim \mathcal{N}(0,1)\,, \quad \tilde{s}_{\mathrm{r}} = G_{z_{\mathrm{r}}^{\mathrm{g}}}^{\mathrm{r}}(E(s))\,. \tag{5}$$

3.4   Local-Global Few-Shot loss (LGFS)

To guide the learning, the resulting image $\tilde{s}$ is compared against the few-shot training set $\mathcal{T}$ with a combination of two loss functions. First, we take inspiration from the state-of-the-art of image style transfer where one image is enough for transferring the *global* appearance of the style scene [12]. Our intuition is that feature statistics alignment, widely used in style transfer, could be less prone to overfitting with respect to adversarial training. Therefore, we align features between $\tilde{s}$ and a target image $t \in \mathcal{T}$ using style loss $\mathcal{L}_{\mathrm{style}}$ as in [12]

$$\mathcal{L}_{\mathrm{style}} = \sum_{k=1}^{K} ||\mu_k(\tilde{s}) - \mu_k(t)||_2 + ||\sigma_k(\tilde{s}) - \sigma_k(t)||_2\,, \tag{6}$$

where (μ_k, σ_k) are the same as in Sec. 3.3. While this is effective in modifying the general image appearance, aligning statistics alone is insufficient to produce realistic outputs. Thus, to provide *local* guidance, *i.e.*, on more fine-grained characteristics, we employ an additional discriminator D_{fs} which is trained to distinguish between rotated patches sampled from $\tilde{s}$ and t. We define the adversarial losses [32]:

$$\begin{aligned} \mathcal{L}_{\mathrm{patch}}^{G} &= ||D_{\mathrm{fs}}(p(\tilde{s})) - 1||_2\,, \\ \mathcal{L}_{\mathrm{patch}}^{D} &= ||D_{\mathrm{fs}}(p(\tilde{s}))||_2 + ||D_{\mathrm{fs}}(p(t)) - 1||_2\,, \end{aligned} \tag{7}$$

where p is a random cropping and rotation function. Note how the *exemplar* residual (from Sec. 3.3) is conditioned on the *same* feature statistics used here—this is what enables the exemplar-based behavior of the network. Also note the interaction between components: backpropagating the LGFS loss *deforms* the manifold learned by multi-target i2i, at the point identified by WMI, thereby injecting $\mathcal{T}$ "between" $\{\mathcal{A}_{\mathrm{id}}, \mathcal{A}_m\}$. We provide additional visualizations of the deformed manifold in the supp. video and material.

3.5   Training strategy

Our framework is fully trained end-to-end and optimizes

$$\min_{\Theta(E,G,G^r,Z),w} \mathcal{L}_{\mathrm{style}} + \mathcal{L}_{\mathrm{patch}}^{G} + \mathcal{L}_{\mathrm{adv}}^{G} \quad \text{and} \min_{\Theta(D_{\mathrm{fs}},D_{\mathrm{mt}})} \mathcal{L}_{\mathrm{patch}}^{D} + \mathcal{L}_{\mathrm{adv}}^{D}\,, \tag{8}$$

where $\Theta(\cdot)$ refers to the network parameters. We train GERM (Sec. 3.3) by randomly selecting one of the exemplar or general mode at each training iteration. For the multi-target settings, we adapt the discriminator and the style encoder of our backbone in a multi-target setup following [5].

4   Experiments

We leverage 4 datasets [46, 45, 6, 43] and 3 translation tasks (Sec. 4.1) and evaluate performances against recent baselines [29, 44, 56, 31, 13] (Sec. 4.2). We further demonstrate the benefit of our few-shot translation on a downstream segmentation task (Sec. 4.3), and rare few-shot scenarios (Sec. 4.4), and finally ablate our contributions (Sec. 4.5). In all, we use MUNIT [13] as our backbone.

4.1   Training setup

Datasets We use four datasets for our experiments.
ACDC   We use ACDC [46] for most of our experiments, using the night/rain/ snow/fog conditions with 400/100/500 images for train/val/test respectively, following official splits. For any individual condition, ACDC also includes geolocalized weakly-paired clear weather day images of same splits.
Dark Zurich   Similar to ACDC, Dark Zurich (DZ) [45] has daytime images paired with nighttime/twilight conditions. Here, we focus on twilight conditions exclusively and use training images from the GOPRO348 sequence only since it exhibits a distinctive twilight appearance. We split the total 819 image pairs into 25/794 for train/test, respectively.
Cityscapes   Cityscapes [6] is used to evaluate ManiFest for training segmentation networks robust to nighttime[1]. It includes 2975/500/1525 annotated images for train/val/test.
VIPER   As anchors, we employ synthetic images from the VIPER dataset [43], using the condition metadata to define splits. 4137/3090/1305/2018/2817 images are extracted from the VIPER training set for day/night/rain/snow/sunset conditions, respectively.

Tasks and evaluation We train our framework on three main tasks:
Day $\mapsto$ Night   on ACDC daytime ($\mathcal{S}$) and nighttime ($\mathcal{T}$).
Clear $\mapsto$ Fog   on ACDC daytime ($\mathcal{S}$) and fog ($\mathcal{T}$).
Day $\mapsto$ Twilight   on DZ daytime ($\mathcal{S}$) and twilight ($\mathcal{T}$).
Unless mentioned otherwise, the (synthetic) anchor domains from VIPER are "night" for Day $\mapsto$ Night and Day $\mapsto$ Twilight, and "day" for Clear $\mapsto$ Fog. We evaluate with the FID [10] and LPIPS [60] metrics. While FID compares feature distance globally, LPIPS compares translated source images and the geolocalized paired image in the target dataset. This is beneficial for evaluating our exemplar modality. For all, we train on downsampled x4 images.

4.2   Comparison with the state-of-the-art

Baselines   We compare with four baselines for few-shot image translation with $|\mathcal{T}| = 25$. We extensively evaluate on the most challenging Day $\mapsto$ Night

---

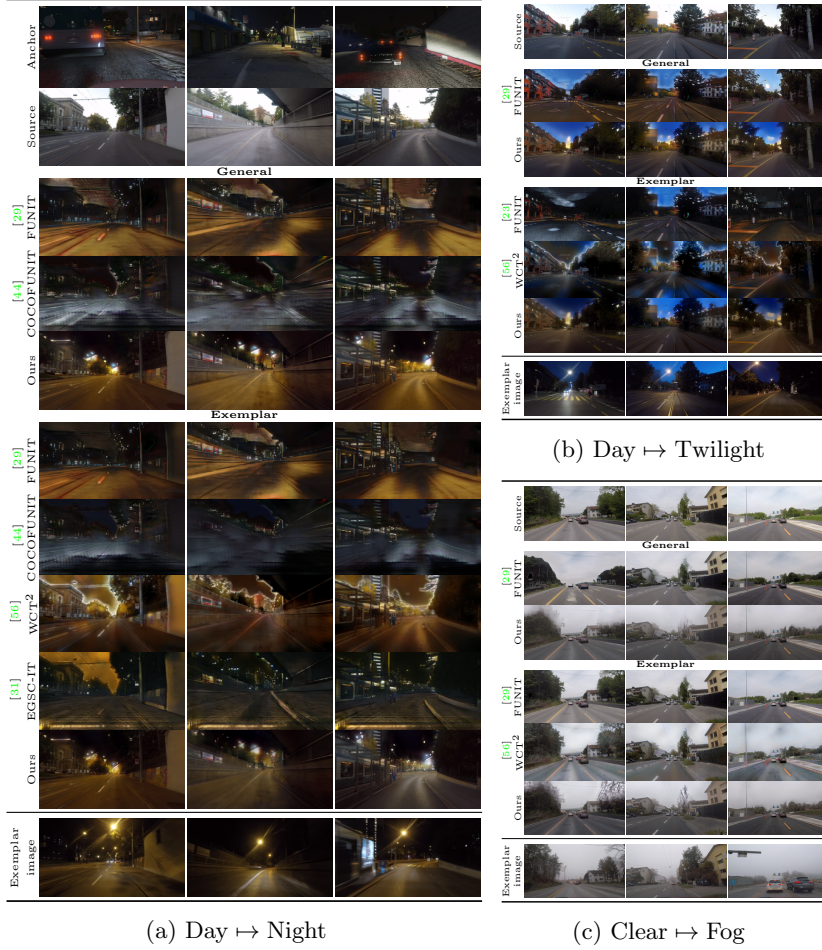[1] ACDC does not provide annotated daytime clear weather sequences.

(a) Day ↦ Night

(b) Day ↦ Twilight

(c) Clear ↦ Fog

Fig. 4: Qualitative evaluation and comparison with the state of the art. We evaluate the a Day ↦ Night, b Day ↦ Twilight, and c Clear ↦ Fog tasks. In all cases, our approach extracts a *general* realistic representation of the few-shot target, and correctly reproduces the style of paired *exemplar* target images. In comparison, existing baselines either has unnecessary similarity with anchors (e.g. FUNIT, COCO-FUNIT, EGSC-IT) or unrealistic artifacts (e.g. WCT^2).

task, and provide insights and comparison for the two others tasks. We evaluate the impact of the few-shot image selection and of $|\mathcal{T}|$ in Sec. 4.5. We compare against the recent FUNIT [29] and COCO-FUNIT [44], trained on $\mathcal{S} \mapsto \mathcal{A}_m$ and adapted following [29, 44] to the few-shot $\mathcal{T}$ (general) or to a single reference image (exemplar). For exemplar image translation, we also add specific baselines. First, we compare with WCT^2 [56], used to transfer the style of the paired target condition to the source one. We also evaluate EGSC-IT [31]. The method is

	Method	$|\mathcal{A}_m|$	$|\mathcal{T}|$	FID↓	LPIPS↓
	MUNIT [13]	0	400	79.20	0.529
	MUNIT [13]	3090	0	132.72	0.613
G	MUNIT [13]	0	25	91.61	0.553
	FUNIT [29]	3090	25	156.97	0.573
	COCO-FUNIT [44]	3090	25	201.67	0.644
	Ours	3090	25	**81.01**	**0.535**
	MUNIT [13]	0	400	87.71	0.522
	MUNIT [13]	3090	0	142.04	0.559
E	MUNIT [13]	0	25	128.73	0.562
	FUNIT [29]	3090	25	136.2	0.572
	COCO-FUNIT [44]	3090	25	193.4	0.646
	EGSC-IT [31]	3090	25	106.68	0.574
	WCT^2 [56]	-	-	105.58	0.580
	Ours	3090	25	**80.57**	**0.525**

(a) Day ↦ Night

	Method	$|\mathcal{A}_m|$	$|\mathcal{T}|$	FID↓	LPIPS↓
G	FUNIT [29]	3090	25	69.53	0.511
	Ours	3090	25	**63.15**	**0.510**
E	FUNIT [29]	3090	25	69.97	0.501
	WCT^2 [56]	-	-	71.77	0.536
	Ours	3090	25	**58.07**	**0.483**

(b) Day ↦ Twilight

	Method	$|\mathcal{A}_m|$	$|\mathcal{T}|$	FID↓	LPIPS↓
G	FUNIT [29]	3090	25	152.90	0.580
	Ours	3090	25	**89.57**	**0.520**
E	FUNIT [29]	3090	25	137.7	0.568
	WCT^2 [56]	-	-	120.9	0.591
	Ours	3090	25	**89.89**	**0.521**

(c) Clear ↦ Fog

Table 1: Quantitative comparison with state of the art. We compare FID and LPIPS on the (a) Day ↦ Night, (b) Day ↦ Twilight and (c) Clear ↦ Fog tasks, for both **G**eneral and **E**xemplar translations. Our approach outperforms all baselines on all tasks, while also being on par (**G**) or even outperforming (**E**) the MUNIT backbone trained on the full dataset for Day ↦ Night in a.

trained by merging $\mathcal{A}_m$ and $\mathcal{T}$ since it should be able to identify inter-domain variability, separating $\mathcal{T}$ styles from $\mathcal{A}_m$ [31]. To define metrics bounds, we also train our MUNIT [13] backbone on $\mathcal{A}_m$, on the full $\mathcal{T}$ set and on $\mathcal{T}$ with $|\mathcal{T}| = 25$. More comparisons with the backbone are in Sec. 4.5. We use the official code provided by the authors for all[2]. More details on baselines are in supp.

Evaluation   We compare qualitative results in Fig. 4. In Day ↦ Night (Fig. 4a), even if the appearance of images in $\mathcal{T}$ is partially transferred on translated images (e.g. road color, darker sky), FUNIT and COCO-FUNIT still retain some characteristics of $\mathbb{A}$ (note, for example, how the street is similar to the GTA one) which worsens the overall image realism.

The same can be observed with EGSC-IT, where the hood of the ego-vehicle in anchor images (first column) is retained and significantly impacts visual results. While WCT^2 exhibits sharp results, it does not correctly map the image context, and it is limited to appearance alignment which leads to artifacts (e.g. yellow sky with white halos). Our method generates significantly better results than the baselines in both the *general* and *exemplar* modalities, with visible differences in all three tasks: the *general* appearance is consistent across test samples, and each result adapts to its *exemplar*. For example, observe how the overall sky colors (Day ↦ Twilight, Fig. 4b) match the exemplar. Here, the exemplars were unseen in training (not part of the few-shot set $\mathcal{T}$), thus GERM generalizes the few-shot learned exemplar behavior. The quantitative evaluation in Tab. 1 is coherent

---

[2] For FUNIT [29] and COCO-FUNIT [44], we modify hyperparameters per authors suggestions to adapt to the ACDC and Dark Zurich datasets.

Model	mIoU % ↑	Acc. % ↑
Baseline (*CS day*)	12.93	45.15
MUNIT [13]	21.22	56.65
Ours	**24.31**	**60.50**
Oracle (*ACDC night*)	49.23	88.47

(a) Quantitative evaluation

(b) Qualitative evaluation

Fig. 5: Segmentation on ACDC-night, for few-shot Day↦Night translations ($|\mathcal{T}| = 25$) (a). We outperform the baseline with noticeably better segmentation in (b) due to the increased quality of our translation.

with the qualitative results, as we always outperform baselines. We perform on par (*general*), or even better (*exemplar*) than the backbone trained on the entire set of 400 training images on Day ↦ Night (Tab. 1a). This result shows that GERM (Sec. 3.3) improves modeling of the exemplar style over AdaIN exemplar style injection [13]. The exemplar behavior may force artifacts following subtle characteristics of the scene (as trees in Fig. 4c), for which the general translation may be advisable.

4.3   Segmentation downstream task

We exploit semantic segmentation to evaluate ManiFest for increasing robustness in challenging scenarios. In Fig. 5 we train HRNet [51] on nighttime versions of Cityscapes [6] obtained by translating the dataset with ManiFest or MUNIT, and evaluating on the ACDC-night validation set labels. We choose the best MUNIT and ManiFest configurations following nighttime realism in Tab. 1a with $|\mathcal{T}| = 25$. As lower and upper bounds we train HRNet either on original Cityscapes (baseline) or on ACDC-night training set (oracle). Fig. 5 shows we outperform the MUNIT backbone (+3.09 mIoU) thanks to our better target domain modeling. Additional results on other domains are in supplementary.

4.4   Rare few-shot scenarios

Few-shot plays its full role with conditions that are rare by nature, difficult or even dangerous to photograph, such as auroras or erupting volcanoes. Fig. 6 shows the capability of ManiFest to learn Mountain↦Volcano or Day↦Aurora, by taking as source and anchor the summer and winter Yosemite dataset [63] splits respectively. Each task uses only 4 images from Google Images as $\mathcal{T}$. We generate realistic erupting volcanoes or auroras starting from mountain images, with contextual understanding (Fig. 6, cols 1–4), where only one mountain is mapped to a volcano and auroras only partially cover the sky. Fig. 6 (cols 5–6) also demonstrate how exemplar characteristics are preserved.

4.5   Ablation studies

Architectural components   We evaluate the contribution of each component in ManiFest (*c.f.* Fig. 2, Sec. 3) using the Day ↦ Night task in the *general*
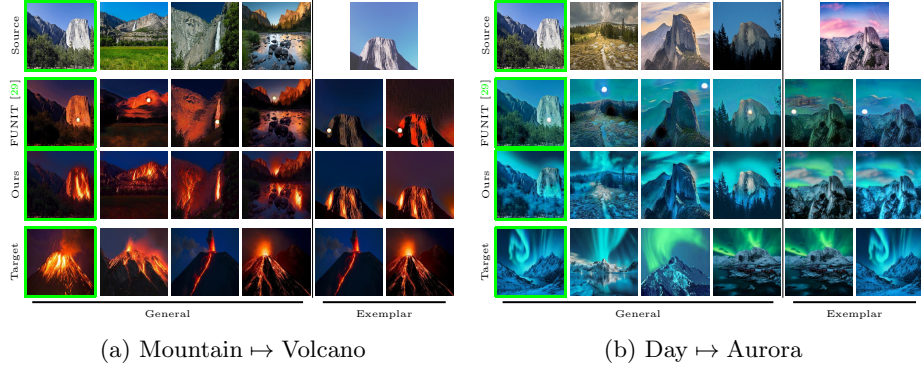
(a) Mountain ↦ Volcano

(b) Day ↦ Aurora

Fig. 6: Qualitative results for the Mountain ↦ Volcano (a) and Day ↦ Aurora (b) tasks. We retain contextual information by only partially mapping mountains to volcanoes and sky to auroras. In the green box we process the same image for ease of comparison. *Exemplar* results show how Ours conforms to Target, effectively reproducing the exemplar image style (cols 5–6). For space reason, we show WCT^2 [56] outputs in Supp.

Component	FID$_\downarrow$	LPIPS$_\downarrow$
w/o $\mathcal{L}_{\text{style}}$	143.66	0.614
w/o $\mathcal{L}_{\text{patch}}$	93.42	0.566
w/o GERM	85.62	0.544
w/o WMI	101.57	0.589
LGFS-only	84.29	0.558
Ours	**81.01**	**0.535**

(a) Quantitative evaluation

(b) Qualitative evaluation

Fig. 7: Ablation study for architectural components. a Removing each component individually lowers quantitative performances, which maps to b decreased visual quality in the generated images.

scenario, and report results in Fig. 7a. The impact of LGFS is studied by removing $\mathcal{L}_{\text{style}}$ or $\mathcal{L}_{\text{patch}}$, showing that both local *and* global guidance are improving translations. Removing the GERM from the training pipeline simultaneously precludes the *exemplar* behavior and worsens the performance, demonstrating the effectiveness of encoding complementary characteristics outside of the manifold spanned by $\mathbb{A}$. The benefit of WMI is evaluated in two experiments. First, the "w/o WMI" setting applies the residual directly on the fake anchor images $\tilde{s}_c$, instead of the interpolated $\tilde{s}_w$ as in Eq. (3). The worse performance relate to synthetic characteristics present in $\tilde{s}_c$ (*e.g.* road texture in Fig. 7b). Second, "LGFS-only" directly uses the LGFS losses in substitution to $\mathcal{L}_{\text{adv}}$, without WMI and GERM components. While it only slightly worsens metrics, the impact on feature consistency is dramatic as shown in Fig. 7b, where the sky presents obvious artifacts and road trivially darkens.

(a) Intra-dataset

$\mathcal{S}$	$\mathcal{T}$	$\mathcal{A}_m$	**FID**$_\downarrow$	**LPIPS**$_\downarrow$
			Day $\mapsto$ Night	
ACDC-Day	ACDC-Night	Day	85.73	0.553
		Night	**81.01**	**0.535**
		Rain	81.38	0.549
		Snow	86.74	0.554
		Sunset	83.83	0.571
		All	83.71	0.547

$\mathcal{S}$	$\mathcal{T}$	$\mathcal{A}_m$	**FID**$_\downarrow$	**LPIPS**$_\downarrow$
			Day $\mapsto$ Twilight	
DZ-Day	DZ-Twilight	Day	64.19	0.505
		Night	63.15	0.510
		Rain	65.33	0.501
		Snow	64.09	0.513
		Sunset	63.78	0.504
		All	**60.98**	**0.469**

$\mathcal{S}$	$\mathcal{T}$	$\mathcal{A}_m$	**FID**$_\downarrow$	**LPIPS**$_\downarrow$
			Clear $\mapsto$ Fog	
ACDC-Clear	ACDC-Fog	Day	**89.57**	**0.520**
		Night	91.79	**0.520**
		Rain	93.15	0.522
		Snow	90.28	0.524
		Sunset	90.11	0.525
		All	92.19	**0.520**

(b) Cross-dataset

$\mathcal{S}$	$\mathcal{T}$	$\mathcal{A}_m$	**FID**$_\downarrow$	**LPIPS**$_\downarrow$
			Day $\mapsto$ Twilight	
ACDC-Day	DZ-Twilight	Day	89.61	*
		Night	90.48	*
		Rain	89.47	*
		Snow	91.49	*
		Sunset	91.77	*
		All	**85.15**	*

Table 2: Study of the impact of anchor domains $\mathbb{A}$ on the $\mathcal{S} \mapsto \mathcal{T}$ translations for intra-dataset (a) and cross-dataset (b) tasks. The stable performance across all tested anchors demonstrates the robustness of our method. For all, we test a multi-anchor setup by using all anchors ("All"). In (b), * means LPIPS cannot be computed due to lack of pairs of matched images (Sec. 4.1).

Anchor selection   We ablate the choice of anchor domain $\mathbb{A}$ by selecting different conditions from the VIPER dataset, namely {Day, Night, Rain, Snow, Sunset}. In particular, we experiment on previous *intra-dataset* ($\mathcal{S}$ and $\mathcal{T}$ taken from the same dataset) tasks, as well as on a *cross-dataset* task in which $\mathcal{S}$ = ACDC-Day and $\mathcal{T}$ = DZ-Twilight. Results in Tab. 2 show how performance remains relatively stable across most anchors. This may seem counter-intuitive since one could, for example, expect that the "Rain" anchor would be a poor choice for the Day $\mapsto$ Night task since rainy and night scenes look different. The results instead show that the WMI only encodes consistency in the transformation, and is thus robust to the choice of anchors. We also test a multi-anchor setup ("All" in Tab. 2), where $\mathbb{A} =\{\mathcal{A}_{\mathrm{id}},$ Day, Night, Rain, Snow, Sunset}. In general, more anchor domains improve performances, ranking either first or second in all cases for at least one metric, due to the additional information available for shaping the manifold in WMI. We hypothesize that multiple anchors helps identifying correspondences between $\mathcal{S}$ and $\mathcal{T}$, benefiting especially the cross-dataset tasks.

Number of images and variability   First we compare our Day $\mapsto$ Night translations against MUNIT [13] for $|\mathcal{T}| = \{25, 20, 15, 10, 5, 1\}$, to understand the effects of few-shot training on the backbone network. Some qualitative *general* outputs are shown in Fig. 8a. While MUNIT overfits and creates unrealistic appearance (25–10 images) or collapse (5, 1 image), we output realistic transformations in all cases, even retaining the image context in the extreme one-shot scenario. This is confirmed by the FID and LPIPS in Figs 8b and 8c for the *general* and *exemplar* scenarios respectively.

   In Tab. 3 we also study variability, evaluating FID and LPIPS for the *general* and *exemplar* cases for $|\mathcal{T}| = \{25, 15, 5, 1\}$ images reporting the results of 7 runs. Overall, the performance remains relatively constant with the exception of the one-shot setup, where despite realistic transfer, the metrics are penalized since the target image itself might not accurately represent the style distribution of the test set.
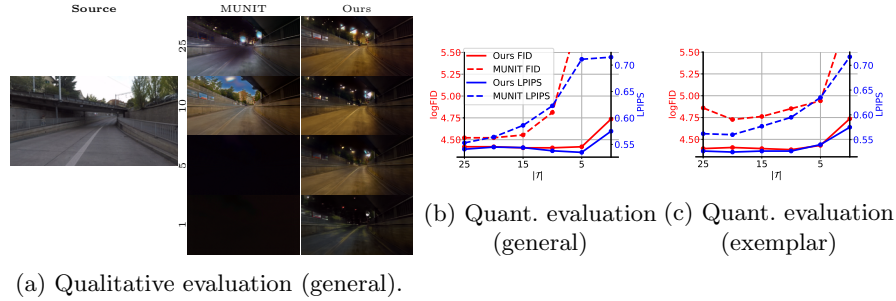
(a) Qualitative evaluation (general).

(b) Quant. evaluation (general)   (c) Quant. evaluation (exemplar)

Fig. 8: Comparison against MUNIT for varying $|\mathcal{T}|$: a qualitatively for the *general* scenario; as well as quantitatively for b *general* and c *exemplar*, always outperforming it.

$|\mathcal{T}|$	$\textbf{FID}_\downarrow$	$\textbf{LPIPS}_\downarrow$
25	82.95 ±2.95	0.541 ±1.85e-2
15	82.21 ±3.09	0.544 ±2.35e-2
5	83.11 ±2.49	0.535 ±2.24e-2
1	114.5 ±34.2	0.575 ±2.37e-2

(a) General

$|\mathcal{T}|$	$\textbf{FID}_\downarrow$	$\textbf{LPIPS}_\downarrow$
25	80.78 ±2.91	0.527 ±0.64e-2
15	80.55 ±2.85	0.527 ±1.07e-2
5	84.40 ±1.88	0.540 ±1.88e-2
1	114.3 ±33.5	0.575 ±2.40e-2

(b) Exemplar

Table 3: Day $\mapsto$ Night ablation on variability by training on 4 few-shot configurations with 7 runs each on general (a) and exemplar (b). $|\mathcal{T}|$ does not impact performance much except for the extreme one-shot scenario, where the network overfits to the seen style. The exemplar behavior performs better due to the style conditioning mechanism.

5   Extensions

5.1   Few-shot continuous manifolds

We investigate the use of ManiFest for performing continuous image translation as in CoMoGAN [40], thus learning the transformation from $\mathcal{S}=day$ to $\mathcal{A}_m=night$ on the Waymo [47] dataset by generating realistically intermediate frontal sun / twilight conditions where we have only few images. Here, we consider *two* few-shot sets ($|\mathcal{T}| = 10$), each one associated to one set of learned weights (w^1, w^2) between identity and night anchors. Results are in Fig. 9, where we also perform comparably to DNI-MUNIT [52] and CoMoGAN which are trained with significantly more intermediate data (4721 vs 20). Please note that estimating w^1 and w^2 reorganizes the transformation realistically (*i.e.* Day $\mapsto$ Frontal sun $\mapsto$ Twilight $\mapsto$ Night) without prior knowledge on the order of few-shot set in the manifold. We evaluate mean rolling FID (mrFID) as in [40] and perform on par or better than baselines (for Model / StarGAN V2 / DNI - CycleGAN / DNI - MUNIT / CoMoGAN / Ours we get 195 / 177 / 155 / 144 / 145 / 145). Only CoMoGAN (mrFID 137) outperforms thanks to its physical guidance.

Fig. 9: By enforcing multiple few-shot manifold deformations, we discover a realistic timelapse appearance comparable with CoMoGAN [40] and DNI-MUNIT [52] (all non few-shot).
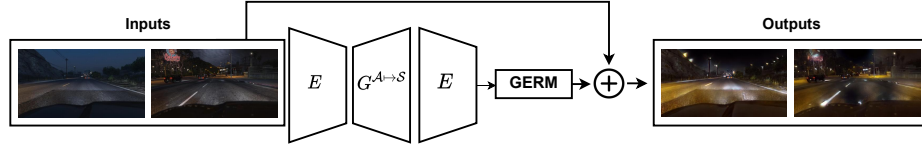

Fig. 10: In the reconstruction cycle $\mathcal{A}_m \mapsto \mathcal{S} \mapsto \mathcal{A}_m$, we can inject the extracted residual with GERM on anchor images to perform an alternative $\mathcal{A}_m \mapsto \mathcal{T}$ transformation.

5.2 Anchor-based translation

The GERM extracts residual information from encoded source images. We investigate the application of residuals on the anchor images themselves, by first translating from $\mathcal{A}_m \mapsto \mathcal{S}$ using our backbone cycle consistency [13], and afterwards re-encoding the fake image in a $\mathcal{S} \mapsto \mathcal{A}_m$ reconstruction *without retraining* (see Fig. 10). This shows how ManiFest simultaneously learns $\mathcal{S} \mapsto \mathcal{T}$ and acceptable $\mathcal{A}_m \mapsto \mathcal{T}$ transformations. The FID w.r.t. ACDC-Night improves from 142 to 130 when applying the residual on the synthetic anchors, thus confirming their shift towards $\mathcal{T}$.

6   Conclusion

In this paper we presented ManiFest, a framework for few-shot i2i which enables translating images to a single *general* style approximating the entire few-shot set (*e.g.*, for photo editing), or reproducing any specific *exemplar* from the set for more variability (*e.g.*, for domain adaptation). We demonstrated its effectiveness outperforming the state-of-the-art on many tasks, ablated its components and provided extensions to the framework.

References

1. Abid, M.A., Hedhli, I., Lalonde, J.F., Gagne, C.: Image-to-image translation with low resolution conditioning. arXiv (2021) 3
2. Cao, J., Hou, L., Yang, M.H., He, R., Sun, Z.: Remix: Towards image-to-image translation with limited data. In: CVPR (2021) 3
3. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: ICLR (2019) 4
4. Cherian, A., Sullivan, A.: Sem-gan: Semantically-consistent image-to-image translation. In: WACV (2019) 3
5. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR (2020) 2, 4, 6
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 7, 10
7. Dell'Eva, A., Pizzati, F., Bertozzi, M., de Charette, R.: Leveraging local domains for image-to-image translation. In: VISAPP (2022) 3
8. Endo, Y., Kanamori, Y.: Few-shot semantic image synthesis using stylegan prior. CoRR (2021) 3
9. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016) 3
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 7
11. Hu, H., Wang, W., Zhou, W., Zhao, W., Li, H.: Model-aware gesture-to-gesture translation. In: CVPR (2021) 3
12. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017) 3, 5, 6
13. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018) 2, 4, 5, 7, 9, 10, 12, 14
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) 2
15. Jia, Z., Yuan, B., Wang, K., Wu, H., Clifford, D., Yuan, Z., Su, H.: Semantically robust unpaired image translation for data with unmatched semantics statistics. In: ICCV (2021) 2
16. Jiang, L., Zhang, C., Huang, M., Liu, C., Shi, J., Loy, C.C.: Tsit: A simple and versatile framework for image-to-image translation. In: ECCV (2020) 2
17. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: NeurIPS (2020) 3
18. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. IJCV (2020) 2
19. Li, P., Yu, X., Yang, Y.: Super-resolving cross-domain face miniatures by peeking at one-shot exemplar. In: ICCV (2021) 1
20. Li, P., Liang, X., Jia, D., Xing, E.P.: Semantic-aware grad-gan for virtual-to-real urban scene adaption. BMVC (2018) 3
21. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: NeurIPS (2017) 3
22. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photorealistic image stylization. In: ECCV (2018) 3

23. Li, Y., Zhang, R., Lu, J., Shechtman, E.: Few-shot image generation with elastic weight consolidation. In: NeurIPS (2020) 3, 8
24. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: CVPR (2019) 1
25. Lin, C.T., Wu, Y.Y., Hsu, P.H., Lai, S.H.: Multimodal structure-consistent image-to-image translation. In: AAAI (2020) 3
26. Lin, J., Wang, Y., He, T., Chen, Z.: Learning to transfer: Unsupervised meta domain translation. In: AAAI (2020) 3
27. Lin, J., Xia, Y., Liu, S., Zhao, S., Chen, Z.: Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. Neurocomputing (2021) 3
28. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NeurIPS (2017) 2
29. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: ICCV (2019) 1, 3, 7, 8, 9, 11
30. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: CVPR (2017) 3
31. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. In: ICLR (2019) 2, 3, 5, 7, 8, 9
32. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV (2017) 6
33. Mo, S., Cho, M., Shin, J.: Instagan: Instance-aware image-to-image translation. ICLR (2019) 3
34. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: CVPR (2021) 2, 3
35. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV (2020) 2
36. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 1
37. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R.: Swapping autoencoder for deep image manipulation. In: NeurIPS (2020) 1
38. Patashnik, O., Danon, D., Zhang, H., Cohen-Or, D.: Balagan: Cross-modal image translation between imbalanced domains. In: CVPR Workshops (2021) 3
39. Pizzati, F., Cerri, P., de Charette, R.: Model-based occlusion disentanglement for image-to-image translation. In: ECCV (2020) 3
40. Pizzati, F., Cerri, P., de Charette, R.: CoMoGAN: continuous model-guided image-to-image translation. In: CVPR (2021) 3, 13, 14
41. Pizzati, F., Cerri, P., de Charette, R.: Guided disentanglement in generative networks. arXiv (2021) 3
42. Ramirez, P.Z., Tonioni, A., Di Stefano, L.: Exploiting semantics in adversarial training for image-level domain adaptation. In: IPAS (2018) 3
43. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV (2017) 7
44. Saito, K., Saenko, K., Liu, M.Y.: COCO-FUNIT: Few-shot unsupervised image translation with a content conditioned style encoder. In: ECCV (2020) 1, 3, 7, 8, 9
45. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. T-PAMI (2020) 1, 7
46. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021) 7

47. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020) 13
48. Tang, H., Xu, D., Yan, Y., Corso, J.J., Torr, P.H., Sebe, N.: Multi-channel attention selection gans for guided image-to-image translation. In: CVPR (2019) 3
49. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global (2010) 3
50. Tremblay, M., Halder, S.S., de Charette, R., Lalonde, J.F.: Rain rendering for evaluating and improving robustness to bad weather. IJCV (2020) 3
51. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. T-PAMI (2019) 10
52. Wang, X., Yu, K., Dong, C., Tang, X., Loy, C.C.: Deep network interpolation for continuous imagery effect transition. In: CVPR (2019) 13, 14
53. Wang, Y., Khan, S., Gonzalez-Garcia, A., Weijer, J.v.d., Khan, F.S.: Semi-supervised learning for few-shot image-to-image translation. In: CVPR (2020) 3
54. Wang, Y., Mantecon, H.L., Lopez-Fuentes, J.v.d.W., Raducanu, B.: Transferi2i: Transfer learning for image-to-image translation from small datasets. In: ICCV (2021) 3
55. Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: CVPR (2019) 3
56. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: ICCV (2019) 3, 7, 8, 9, 11
57. Yu, X., Chen, Y., Liu, S., Li, T., Li, G.: Multi-mapping image-to-image translation via learning disentanglement. In: NeurIPS (2019) 2
58. Zhan, F., Yu, Y., Cui, K., Zhang, G., Lu, S., Pan, J., Zhang, C., Ma, F., Xie, X., Miao, C.: Unbalanced feature transport for exemplar-based image translation. In: CVPR (2021) 3
59. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: CVPR (2020) 3
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 7
61. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. NeurIPS (2020) 3
62. Zheng, C., Cham, T.J., Cai, J.: The spatially-correlative loss for various image translation tasks. In: CVPR (2021) 2
63. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: CVPR (2017) 2, 10
64. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NeurIPS (2017) 2
65. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: CVPR (2020) 3
66. Zhu, Z., Xu, Z., You, A., Bai, X.: Semantically multi-modal image synthesis. In: CVPR (2020) 3