Leveraging Local Domains for Image-to-Image Translation

Anthony Dell'Eva[1], Fabio Pizzati[1,2], Massimo Bertozzi[3] and Raoul de Charette[2]
[1] *VisLab, Parma, Italy*
[2] *Inria, Paris, France*
[3] *University of Parma, Parma, Italy*

Keywords: Computer Vision, Generative Networks, Few-shot Learning, Autonomous Driving, Lane Detection, Segmentation.

Abstract: Image-to-image (i2i) networks struggle to capture local changes because they do not affect the global scene structure. For example, translating from highway scenes to offroad, i2i networks easily focus on global color features but ignore obvious traits for humans like the absence of lane markings. In this paper, we leverage human knowledge about spatial domain characteristics which we refer to as 'local domains' and demonstrate its benefit for image-to-image translation. Relying on a simple geometrical guidance, we train a patch-based GAN on few source data and hallucinate a new unseen domain which subsequently eases transfer learning to target. We experiment on three tasks ranging from unstructured environments to adverse weather. Our comprehensive evaluation setting shows we are able to generate realistic translations, with minimal priors, and training only on a few images. Furthermore, when trained on our translations images we show that all tested proxy tasks are significantly improved, without ever seeing target domain at training.

1 INTRODUCTION

Apart from their appealing translations, image-to-image (i2i) GAN networks also offer an alternative to the supervised-learning paradigm. Indeed, as translations share features characteristics with the target domain they can be used to fine-tune proxy tasks, reducing the need for target annotations. However, i2i GANs perform well at learning global scene changes – winter↦summer, paints, etc. (Liu et al., 2017; Zhu et al., 2017a), – but struggle to learn subtle local changes. Instead, we leverage human domain knowledge to guide i2i and improve proxy tasks on target, *without seeing target images*. This is of paramount importance for real-world applications like autonomous driving (Schutera et al., 2020; Bruls et al., 2019; Romera et al., 2019) which must operate safely in all hazardous conditions – some of which are rarely observed.

We propose a method exploiting human knowledge about source and target, to identify domain-specific local characteristics which we call *local domains* (Fig. 1, top). The latter are used as guidance to perform patches translations on *source only*, thus hallucinating a new unseen domain.

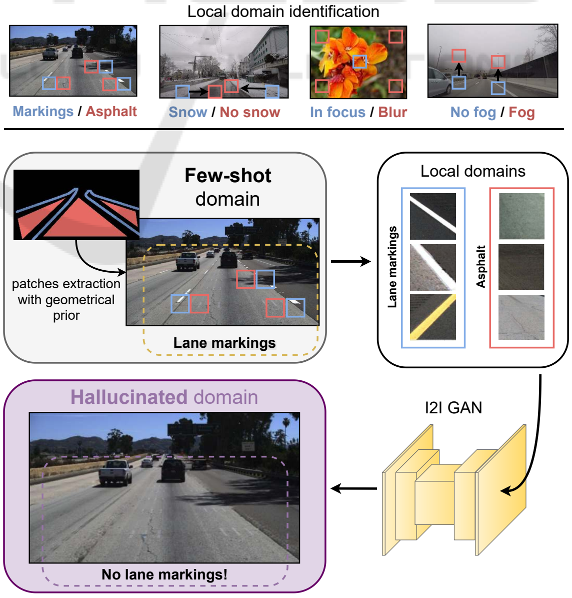


Figure 1: **Overview.** Our method is able to generate images of unseen domains, leveraging geometrical-guidance to extract patches of *local domains*, i.e. spatially defined subdomains, on source images. Here, we generate an image without any lane markings training only on an extremely small amount of images with well-defined lane markings.

Dell'Eva, A., Pizzati, F., Bertozzi, M. and de Charette, R.
Leveraging Local Domains for Image-to-Image Translation.
DOI: 10.5220/0010848900003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 179-189
ISBN: 978-989-758-555-5; ISSN: 2184-4321

An example in Fig. 1 bottom shows we leverage local domains knowledge about ‘lane markings’ and ‘asphalt’ to hallucinate a new domain without lane markings. The latter domain can then be used to increase robustness on unstructured road environments which are typically hard to capture but may cause dramatic failures.

Experimental evidence in this paper shows indeed that our new domain acts as a bridge leading to a performance boost on target. Notably, our method exhibits few-shot capabilities, requiring only source images and minimal human knowledge about the target. In short, the main contributions of this paper are:

- we introduce and define *local domains* as being domain-specific spatial characteristics (Sec. 3.1),
- to the best of our knowledge, we propose the first geometrical-guided patch-based i2i, leveraging our *local domains* priors (Sec. 3.2) and enabling continuous geometrical translation (Sec. 3.3),
- we experiment on three different tasks in a few-shot setting, showing that our translations lead to better performance on all target proxy tasks (Sec. 4).

2 RELATED WORKS

Image-to-Image Translation (i2i). In 2017, i2i was introduced as an application of conditional GANs in (Isola et al., 2017), extended by (Liu et al., 2017; Zhu et al., 2017b) for multi-modality or better performances. In more recent approaches, the obvious limitation of requiring paired images for training has been removed (Huang et al., 2018b; Zhu et al., 2017a; Lee et al., 2020; Yi et al., 2017). Recently, there has been an emergence of attention-based (Mejjati et al., 2018; Ma et al., 2018; Tang et al., 2019a; Kim et al., 2020; Lin et al., 2021b) or memory-based (Jeong et al., 2021) methods, which further guarantee more realism or increased network capabilities. Some methods guarantee multi-domain translations (Choi et al., 2018; Choi et al., 2020). The first approach efficiently exploiting a patch-wise decomposition of images has been CUT (Park et al., 2020), which exploits patches from different domains to impose contrastive learning constraints. All these methods use different data as source and target and are unable to identify inter-domain transformation by default.

Image Translation with Less Supervision. A recent field of study focuses on reducing the number of images necessary for training i2i networks. For instance, in BalaGAN (Patashnik et al., 2021) they exploit domain clustering in order to make the training robust to classes with few examples. Other strategies use self-supervision (Wang et al., 2020) or latent space interpolations (Cao et al., 2021) in order to avoid the discriminator overfitting and train on extremely small datasets. Differently, FUNIT (Liu et al., 2019) and COCO-FUNIT (Saito et al., 2020) generalize to few-shot domains at inference stage. Some other works try to work with less supervision at the domain level, on a mixed target domain (Pizzati et al., 2021a) or without even source and target domain distinctions (Baek et al., 2020; Lee et al., 2021). It is worth noticing that some methods are trained on single images, as SinGAN (Shaham et al., 2019), employable for image editing tasks. Finally, ZstGAN (Lin et al., 2021a) exploits textual inputs for zero-shot image translation.

Prior-guided Image Translation. Several priors can be exploited to increase image translation effectiveness, with several degrees of supervision as bounding boxes (Shen et al., 2019; Bhattacharjee et al., 2020), semantic maps (Li et al., 2018; Ramirez et al., 2018; Tang et al., 2019b; Cherian and Sullivan, 2019; Zhu et al., 2020b; Zhu et al., 2020a; Lin et al., 2020; Ma et al., 2019; Park et al., 2019) or instance labels (Mo et al., 2019; Xu et al., 2021). Another line of works exploits physical models as priors for translation enhancement (Halder et al., 2019; Tremblay et al., 2020), disentanglement (Pizzati et al., 2021b), or guidance (Pizzati et al., 2021a). Importantly, scene geometry could be used as a prior, with learned correspondences (Wu et al., 2019) or by exploiting additional modalities (Arar et al., 2020). Some use text for image editing purposes (Liu et al., 2020). Others exploit full semantic maps for road randomization (Bruls et al., 2019), to generalize across challenging lane detection scenarios. However, they are limited to annotated road layouts and constrained by expensive complete segmentation maps.

3 METHOD

We address the problem of image to image translation accounting for *source* and *target* domains having predominant local transformations. As such, leveraging *only source* data, our proposal hallucinates a new *unseen* intermediate domain which can be used to ease transfer learning towards *target*. An overview of our pipeline is in Fig. 2.

In the following, we introduce our definition of local domains (Sec. 3.1) and propose a geometrical-guided patch-based strategy to learn translation be-

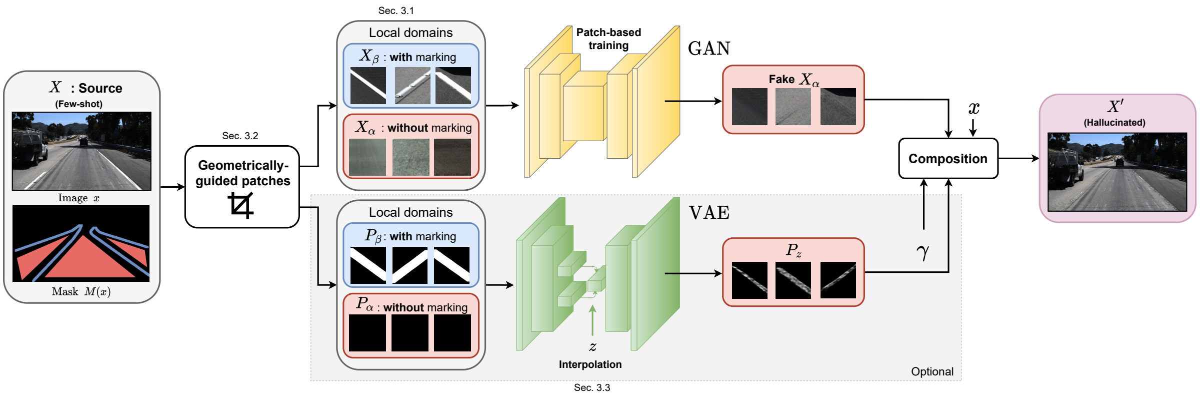
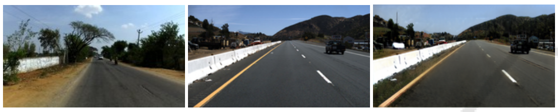


Figure 2: **Architecture pipeline.** Our method exploits knowledge about *local domains* (Sec. 3.1) and relies on geometrical-prior to extract samples of local domains in *source only* (Sec. 3.2) that train a patch-based GAN. Here, source is having "lane markings" and "asphalt" local domains (X_α and X_β, respectively) while target have only "asphalt" (X_β), learning $X_\alpha \mapsto X_\beta$ further reduces the gap with target. An optional local domain interpolation strategy (Sec. 3.3) is added for generating geometrically continuous translation between local domains (here, simulating lane degradation).



IDD (Varma et al., 2019) | TuSimple (TuSimple, 2021) | TuSimple↦IDD

Figure 3: **Translation with CycleGAN (Zhu et al., 2017a).** Sample output shows that i2i is prone to transfer global features (here, sky color) but neglects evident local features for humans as the street structure (note that IDD has no lane markings).

tween the latter (Sec. 3.2). For some local domains, we also show that a continuous geometrical translation can be learned from the interpolation of a mask (Sec. 3.3). Finally, we describe our training strategy showing few shot capabilities (Sec. 3.4).

3.1 Local Domains

Image-to-image (i2i) networks learn a mapping function $G : X \mapsto Y$ from a source domain X to a target domain Y, such that the distribution $P_{G(X)}$ approximates P_Y. The goal is to transfer the features of domain Y to samples from X while preserving their content. This works well for transformation globally affecting the scene (eg. summer to winter) but struggles to capture the mappings of local changes due to the under-constrained settings of the system. A simple failure example, shown in Fig. 3, is the translation from outdoor images having lane markings, to images having no (or degraded) lane markings. As it seeks global changes, the i2i is likely to transfer unintended characteristics while missing the subtle – but consistent – local changes (here, the lane markings).

To overcome this, we introduce *local domains* which are sub-domains *spatially defined* – for example, lane markings, asphalt, etc. Formally, we define domain X as the composition of local domains, denoted $\{X_\alpha, ..., X_\omega\}$, and the remaining sub-domains written X_O. Considering only two local domains of interest, it writes:

$$X = \{X_O, X_\alpha, X_\beta\}. \tag{1}$$

Because we consider only source and target domains sharing at least one local domain, say $_\alpha$, we write Y as:

$$Y = \{Y_O, Y_\alpha\}, \tag{2}$$

so that the Kullback-Leibler divergence $\text{KL}(X_\alpha, Y_\alpha)$ is close to 0. Instead of learning the direct mapping of $X \mapsto Y$, we propose to learn local domain mappings, such as $X_\beta \mapsto X_\alpha$. If such mapping is applied systematically on all samples from X, we get a new domain X' without $_\beta$, so:

$$X' = \{X_O, X_\alpha\}, \tag{3}$$

where domain X' is unseen and thus hallucinated. Considering that X' and Y share the same local domains, they are subsequently closer: $\text{KL}(Y, X') < \text{KL}(Y, X)$.

Our intuition is that when training target data is hard to get, our hallucinated domain X' can ease transfer learning. Notably here, our method only requires a priori knowledge of the shared local domains in source and target.

3.2 Geometrically-guided Patches

Learning the mapping between local domains requires extracting local domain samples. To do so we leverage patches corresponding to either local domains in the source dataset only. We rely here on a simple geometrical guidance from a mask $M(.)$ to extract random patches centered around a given local domain.

Considering x an image in source domain X, we extract $\mathcal{X}_\alpha$ the unordered set of patches of fixed dimension, so that:

$$\mathcal{X}_\alpha = \{\{x_{p_0}, x_{p_1}, \dots, x_{p_m} | p \in M_\alpha(x)\} | \forall x \in X\}, \quad (4)$$

having m the number of patches per image, and $M_\alpha(x) = [\![M(x) = \alpha]\!]$ with $[\![.]\!]$ the Iverson brackets. Literally, $M(x)$ is our geometrical prior – a 2D mask of the same size as x – encoding the position of local domains. Subsequently, $M_\alpha(x)$ is filled with ones where local domain X_α is and zeros elsewhere. Similarly to Eq. 4, we extract the set X_β from $M_\beta(x)$ and X.

In practice, the geometrical prior $M(x)$ is often simply derivable from the image labels. For example, the position of lane marking and asphalt can both be extracted from image labels. In some cases, the position of local domains is constant dataset-wise and we use a fixed geometrical prior, so $M(x) = M$. This is for example the case for portraits datasets, where faces are likely to be centered and background located along the image edges.

Having collected the two sets of patches $\mathcal{X}_\alpha$ and $\mathcal{X}_\beta$, a straightforward patch-based GAN can learn $\mathcal{X}_\alpha \mapsto \mathcal{X}_\beta$. In some cases, $\mathcal{X}_\alpha$ and $\mathcal{X}_\beta$ being of similar nature we demonstrate spatial interpolation is beneficial.

3.3 Local Domains Interpolation

Continuous i2i are extensively studied (Gong et al., 2019; Wang et al., 2019; Lample et al., 2017), but existing methods are not suitable for translation affecting only local regions as in our problem setting (see Sec. 4). Instead, we learn a non-linear geometrical interpolation of patch masks, leveraging a variational autoencoder (VAE).

Previously we described each patch as encompassing a single local domain but, in reality, patches often mix multiple local domains. This is the case of lane markings patches, shown in Fig. 2, that contain asphalt too. Hence, along with the set of local domains patches we extract the sets $\mathcal{P}_\alpha$ and $\mathcal{P}_\beta$ directly from our geometrical guidance $M(.)$, and seek to continuously interpolate $P_\alpha \mapsto P_\beta$.

In practice, our VAE having encoder $E(.)$ and decoder $D(.)$ is trained in the standard fashion, but at inference it yields the latent representation h_Z corresponding to the linear combination of $E(p_\alpha)$ and $E(p_\beta)$, having $p_\alpha \in \mathcal{P}_\alpha$ and $p_\beta \in \mathcal{P}_\beta$, respectively[1]. Formally:

$$\begin{aligned} h_Z = E(p_\alpha)\, z + E(p_\beta)\,(1-z), \\ p_z = D(h_Z), \end{aligned} \quad (5)$$

where $z \in [0,1]$ encodes the progress along $P_\alpha \mapsto P_\beta$. The final interpolated patch x_z is the composite between x_α and x_β patches, following the VAE output. It writes:

$$\begin{aligned} x_z = x_\alpha m + x_\beta\,(1-m)\,, \\ \text{with } m = \gamma p_z\,, \end{aligned} \quad (6)$$

$\gamma \in [0,1]$ being an arbitrary controlled blending parameter adding a degree of freedom to our model. Furthermore, notice that the stochastic VAE behavior further increases variability, beneficial for proxy tasks.

3.4 Training

We train our pipeline, the patch-based GAN and the optional VAE, leveraging only images from the source domain and geometrical priors about local domains. The patch-based GAN is trained on $X_\alpha \mapsto X_\beta$ (Sec. 3.2) minimizing the LSGAN (Mao et al., 2017) adversarial loss:

$$\begin{aligned} y_f &= G(x), \\ \mathcal{L}_G(y_f) &= \mathbb{E}_{x \sim P_X(x)}\left[(D(y_f)-1)^2\right], \\ \mathcal{L}_D(y_f, y) &= \mathbb{E}_{x \sim P_X(x)}\left[(D(y_f))^2\right] + \\ &\quad + \mathbb{E}_{y \sim P_Y(y)}\left[(D(y)-1)^2\right], \end{aligned} \quad (7)$$

along with task-specific losses. If used, the VAE interpolation (Sec. 3.3) is trained with standard ELBO strategy (Blei et al., 2017), minimizing reconstruction loss along with a regularizer:

$$\begin{aligned} \mathcal{L}_{VAE} = -\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) + \\ + D_{KL}(q_\phi(z|x)||p(z)). \end{aligned} \quad (8)$$

At inference time, the full image is fed to the GAN backbone to produce the translated image, while the corresponding full interpolation mask is obtained processing mask patches independently and then stitching them together with a simple algorithm. Of note, our method has important few-shot capabilities. As we train only on source patches, a reduced number of image samples is sufficient to get reasonable data diversity, which we further demonstrated in the following section.

4 EXPERIMENTS

We evaluate our method on 3 different tasks, namely lane markings degradation, snow addition and deblurring, leveraging 5 recent datasets (TuSimple, 2021;

[1] Our formalism includes VAE reparametrization in $E(.)$

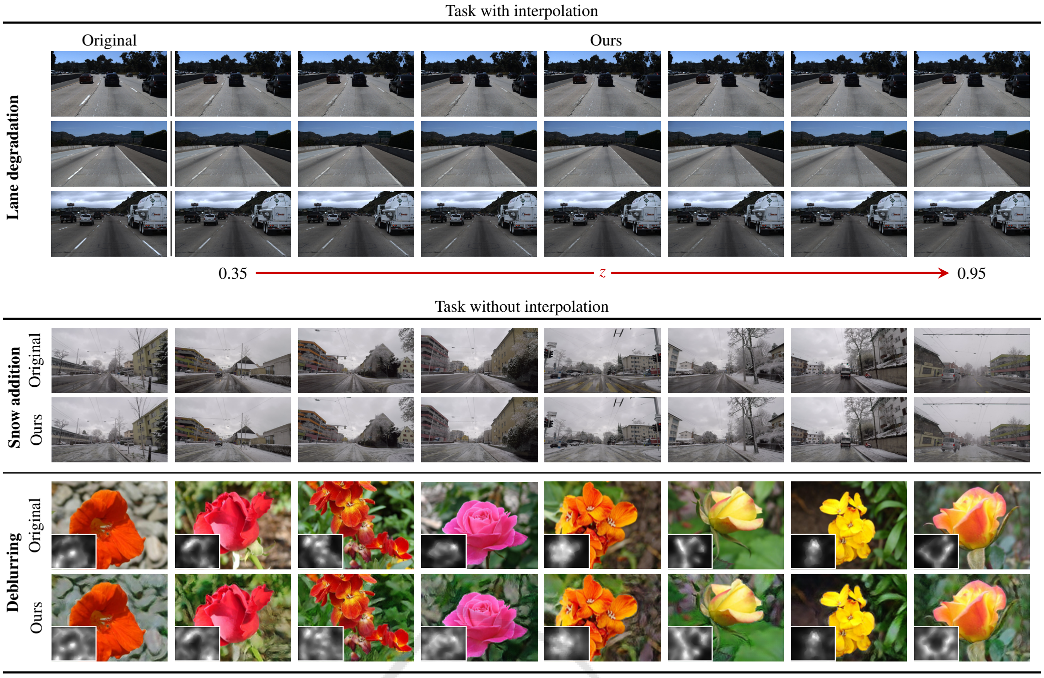


Figure 4: **Qualitative results.** For each task we show the original input image and our output with the $X_\beta \mapsto X_\alpha$ local domains translation. **Lane degradation**: sample translations on TuSimple (TuSimple, 2021) test set with increasing degradation $z \in [0.35, 0.95]$ from left to right, blending variable $\gamma = 0.75$. **Snow addition**: augmentation of ACDC (Sakaridis et al., 2021) validation set, only road is involved in the transformation. **Deblurring**: Original flower images (Nilsback and Zisserman, 2008) and our deblurred version. Bottom left insets show the in-focus map (Golestaneh and Karam, 2017) which are whiter on average (ie. less blurred) for ours.

Varma et al., 2019; Sakaridis et al., 2021; Cordts et al., 2016; Nilsback and Zisserman, 2008), and evaluating our translation both against i2i baselines and on proxy tasks. In Sec. 4.1 we provide details on our tasks, while Secs. 4.2, 4.3 report extensive qualitative and quantitative evaluation.

4.1 Tasks Definitions

We describe our three task below, detailing the learned local domains translation $X_\beta \mapsto X_\alpha$.

Lane Degradation. Here, we leverage the highway TuSimple (TuSimple, 2021) dataset having clear lane markings. For local domains, we chose *lane marking* (X_β) and *asphalt* (X_α) exploiting geometrical priors from the provided lane labels and assuming near by asphalt. We use our interpolation strategy (Sec. 3.3) accounting for both degradation and blending. Importantly, we train only on 15 images (1280x720) to demonstrate few-shot capabilities, with 30 patches per image of size 128x128, 200x200 and 256x256. Backbones are DeepFillv2 (Yu et al., 2019) as GAN and IntroVAE (Huang et al., 2018a) for interpolation. The latter is trained with a binarized difference mask from lane inpainting and original image. We evaluate our translations on the standard 358/2782 val/test sets of TuSimple. In addition to demonstrating generalization, we evaluate several lane detectors on 110 images from the India Driving Dataset (IDD) (Varma et al., 2019) – never seen during training – having degraded lane markings which we manually annotated.

Snow Addition. Here, we rely on snowy images from the recent Adverse Driving Conditions Dataset (ACDC) (Sakaridis et al., 2021), which typically have snow only on the sidewalk and *not* on the road. The task is to add snow on the road. Logically, local domains are *road* (X_β) and snowy *sidewalk* (X_α), exploiting semantic labels as priors. Again, we train only with 15 images with 30 patches (128x128) per image , using CycleGAN (Zhu et al., 2017a) with default hyperparameters. No interpolation is used.

We evaluate on the original val/test set of ACDC having 100/500 images. To increase generalization for the segmentation task in snowy weather, we also augment Cityscapes (Cordts et al., 2016) with the same trained network.

Table 1: **Deblurring performance.** Average of the in-focus maps (Golestaneh and Karam, 2017) on the Oxford Flowers (Nilsback and Zisserman, 2008) test set shows our method efficiently deblur the input images despite a trivial dataset-wise geometrical prior.

Images	In-focus avg↑
Original	1.28
Ours *(deblurred)*	**1.53**

Deblurring. We leverage the Oxford 102 Flower dataset (Nilsback and Zisserman, 2008) to learn turning shallow Depth of Field (DoF) photos to deep DoF, therefore seeking to deblur the image. As blur is not labeled, we rely on a simple *dataset-wise* geometrical prior, i.e. that the image center is always in-focus and image corners are always out-of-focus. Local domains are *out-of-focus* (X_β) and *in-focus* (X_α). Since we use only 8 patches per image (4 in-focus, 4 out-of-focus, 128x128), we train our CycleGAN (Zhu et al., 2017a) with 400 images, adding a task-specific objective function defined as the composition of a color consistency loss and an in-focus loss:

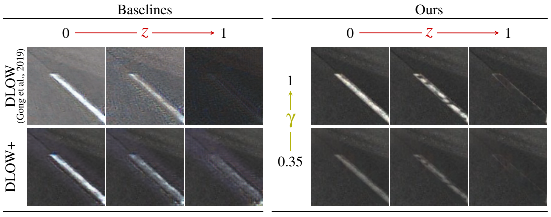
$$\mathcal{L}_{deblur} = D_{KL}(H[x]||H[G(x)]) + \frac{1}{\sigma^2_{LoG(G(x))}}, \qquad (9)$$

with $H[.]$ the image histogram and $\sigma^2_{LoG(.)}$ the Laplacian of Gaussian variance. A color jitter augmentation is applied to ensure discriminator invariance to color. We do not use interpolation. At inference, we exclude foreground since it impacts translation quality due to the identity loss in CycleGAN (Zhu et al., 2017a).

4.2 Evaluation

4.2.1 Translation Quality

Qualitative results are visible in Fig. 4 and show our method outputs realistic translations for all tasks. In detail, we are able to modify lanes (first three rows) on TuSimple with different degrees of degradation (from left to right). On snow addition, images show plausible snow on ACDC roads (middle two rows), preserving shadows. Finally, on our deblurring task (bottom two rows) the flowers background appears in-focus, exhibiting sharper edges. To better size the benefit on this last task, flower images have as inset the in-focus map computed with (Golestaneh and Karam, 2017). In extenso, white means in-focus.



(a) Qualitative

Network	FID↓	LPIPS↓
DLOW (Gong et al., 2019)	211.7	0.4942
DLOW+	155.6	0.4206
Ours w/o blending	154.7	0.3434
Ours	**135.4**	**0.3254**

(b) GAN metrics

Figure 5: **Lane translations.** (a) Qualitative comparison of lane degradation on patches with baselines. Our method is the only one to output a realistic degradation. (b) GAN metrics on the lane degradation task prove the benefit of our method.

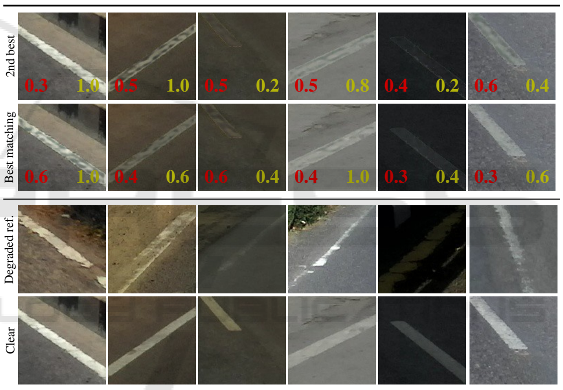


Figure 6: **Evaluation of lane degradation on patches taken from IDD dataset (Varma et al., 2019)**. We associate clear patches (bottom row) to degraded ones (third row) by minimizing LPIPS. Applying our method to clean images variating the z and γ parameters (shown in the images), we subsequently lower the LPIPS. We display the best and second-best translation in terms of LPIPS. The similarities of our results with the degraded patches prove the efficacy of our LPIPS-based evaluation.

Of note, evaluating GAN metrics on target images would be biased since we use only source images – unlike existing i2i –. They are reported in Appendix for the sake of completeness. To provide a quantitative quality evaluation, Tab. 1 reports the in-focus average proving our translations are significantly more in-focus (+0.24) than original images.

4.2.2 Interpolation Quality

For the lane degradation task, we compare our interpolations against the continuous i2i DLOW (Gong et al., 2019) baseline, trained on the same data. As it suffers from evident color artifacts, we introduce DLOW+: a custom version using lane mask as additional channel input, masked reconstruction loss, and masked input-output blending. For DLOW/DLOW+,

Table 2: **Lane detection on TuSimple and IDD.** Performance of lane detectors when trained on TuSimple source (*none*) or our degraded translations (*ours*). The latter significantly outperforms baseline, while retaining equivalent performances on TuSimple images.

Detector	Translation	TuSimple (TuSimple, 2021)			IDD (Varma et al., 2019)		
		Acc. ↑	FP ↓	FN ↓	Acc. ↑	FP ↓	FN ↓
SCNN (Pan et al., 2017)	none (source)	**0.946**	**0.052**	**0.069**	0.617	0.538	0.741
	Ours	0.945	0.058	0.072	**0.730**	**0.453**	**0.577**
RESA (Zheng et al., 2021)	none (source)	**0.952**	**0.056**	**0.065**	0.639	0.720	0.800
	Ours	0.951	0.059	0.068	**0.671**	**0.686**	**0.761**

we regulate the walk on the discovered manifold of each network with a *domainness* variable z – which amounts to our lane degradation.

With respect to baselines, it is visible in Fig. 5a that our degraded lane translations are more realistic for all z since DLOW and DLOW+ discover simpler transformations, just regulating color homogeneously.

For quantification, we compare translations against real degraded lane markings from IDD and report FID and LPIPS in Fig. 5b. In detail, we select 35/62 clear/degraded lane patches from IDD test set, and couple those with minimum LPIPS (Zhang et al., 2018) distance. Intuitively, we pair similar clear and degraded lane markings together. Pairs are shown in the two bottom rows of Fig. 6. We then degrade each clear image with ours / DLOW / DLOW+, generating several degraded versions, and use the best degrading version in terms of LPIPS w.r.t. its clear match to compute GAN metrics. Fig. 5b shows we outperform baseline on both metrics significantly (roughly, -20 FID, and -0.1 LPIPS), demonstrating the realism of our lane degradation. Since baselines are not using any explicit blending as us (see γ in Eq. 6), we also evaluate “ours w/o blending” using $m = p_z$ in Eq. 6, which still outperforms baselines.

4.2.3 Proxy Tasks

Here, we study the applicability of our pipeline to increase the robustness of existing lane detection and semantic segmentation networks.

Lane Detection. We aim here to make lane detectors robust to *unseen* degraded lane markings. To do so, we train two state-of-the-art detectors, SCNN (Pan et al., 2017) and RESA (Zheng et al., 2021), on both TuSimple original images and our translated version (mixing with 5% probability and randomizing z and γ). The models are tested on both the TuSimple test set and our 110 labeled IDD images, the latter having severely degraded lane markings.

From the quantitative results in Tab. 2, we observe that with our source degraded translations both detectors severely outperform the baselines using clear source on the challenging IDD, while maintaining on-par performances on TuSimple with clear markings. In particular, for SCNN we improve by +11.3% the accuracy, −8.5% the false positives and −16.4% the false negatives. Sample qualitative results are in Fig. 7 and showcase the robustness of our method on degraded or even absent street lines. We conjecture that our degraded translations forced the network to rely on stronger contextual information.

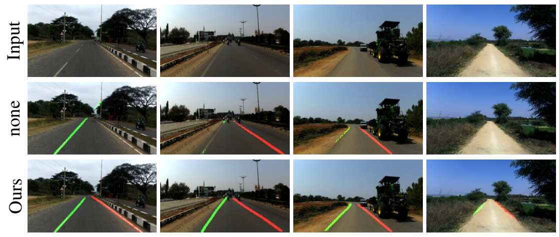


Figure 7: **SCNN (Pan et al., 2017) lane detection on IDD (Varma et al., 2019).** Training on generated images with degraded lanes makes existing lane detectors – such as SCNN (Pan et al., 2017) – resistant to scenes with damaged (first three columns) or no (last column) lane markings.

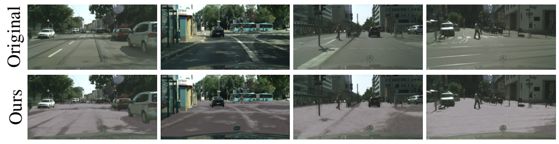


Figure 8: **Cityscapes images with snow added**. We add snow on roads and sidewalks of the Cityscapes training set to train semantic segmentation networks robust to snow. Cityscapes exhibits a domain shift with respect to ACDC, but our method is still able to generate acceptable snow.

Semantic Segmentation. Here, we seek to improve segmentation in snowy driving conditions. We train three state-of-the-art semantic segmentation models, namely DeepLabv3+ (Chen et al., 2018), PSANet (Zhao et al., 2018) and OCRNet (Yuan et al., 2020), with either clear Cityscapes images and snowy Cityscapes images translated with our method. Translation examples are available in Fig. 8, where we add snow on sidewalks and roads by using Cityscapes se-

Table 3: **Semantic segmentation on ACDC (Sakaridis et al., 2021) snow.** We train multiple segmentation networks on Cityscapes (Cord and Aubert, 2011) with added snow with our method and test on ACDC (Sakaridis et al., 2021) snow validation, consistently improving generalization capabilities.

Model	Translations	road IoU ↑	sidewalk IoU ↑	mIoU ↑
DeepLabv3+ (Chen et al., 2018)	none (source)	74.95	39.52	45.31
	Ours	**80.56**	**49.52**	**47.64**
PSANet (Zhao et al., 2018)	none (source)	**74.29**	30.71	42.97
	Ours	74.01	**36.28**	**43.85**
OCRNet (Yuan et al., 2020)	none (source)	82.30	45.60	54.54
	Ours	**82.78**	**54.69**	**55.48**

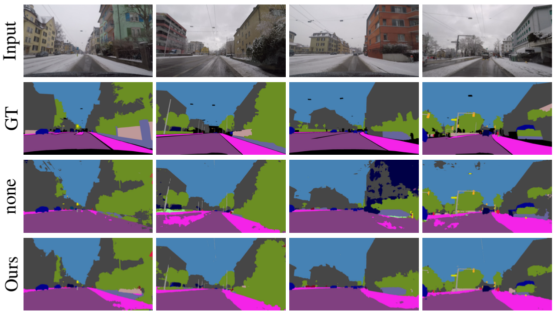


Figure 9: **DeepLabv3+ (Chen et al., 2018) on ACDC (Sakaridis et al., 2021) snow.** Training with our generated images brings improvements in segmentation of snowy scenes in ACDC (Sakaridis et al., 2021), especially in the road and sidewalk classes.

Table 4: **Data ablation on TuSimple.** The use of data on the lane degradation task (TuSimple ↦ IDD) is ablated by varying the number of images and patches per image in the training set, and evaluating GAN similarity metrics (see Sec. 4.2.2) on IDD.

TuSimple samples (%)	Patches/img	LPIPS↓	FID↓
15 (0.4%)	1	0.3296	148.22
15 (0.4%)	5	0.3295	135.53
15 (0.4%)	30	0.3254	131.73
15 (0.4%)	60	0.3246	127.94
15 (0.4%)	150	**0.3222**	**126.94**
50 (1.4%)	30	0.3236	129.42
150 (4.1%)	30	0.3221	**124.79**
500 (14%)	30	0.3234	128.56
3626 (100%)	30	**0.3218**	125.56

mantic maps. Visual results remain acceptable and snow is added uniformly on both semantic classes, even if inference on Cityscapes brings a consistent domain shift with respect to training patches on ACDC. In detail, for the latter we augment images with 10% (DeepLabv3+, PSANet) or 5% (OCRNet) probability. The models are evaluated on the ACDC snow validation set.

Tab. 3 shows the benefit of our augmented images (*Ours*) to consistently improve the performance on road or sidewalk (our two local domains) and mean IoU for all networks. From Fig. 9 it is visible that the model trained with our augmentation strategy is able to better detect roads and footpaths in difficult weather conditions with respect to the baseline, which is not capable of properly discriminating between them if they are covered with snow.

4.3 Ablation Study

Training Images and Patches. As mentioned our method requires very few images to train. Here, we study the effect of number of images and patches per image on the lane degradation task. To measure its impact, we use LPIPS (Zhang et al., 2018) and FID (Heusel et al., 2017) following Sec. 4.2.2.

Results in Tab. 4 show, as expected, better translation with the increase of both the number of images and the number of patches extracted per each image. However, we also denote the few-shot capability of our method and the minimal benefit of using a large number of images.

Table 5: **Augmentation percentage ablation on Cityscapes.** The effectiveness of our snow addition translation is ablated by varying the probability of Cityscapes augmented images shown to DeepLabv3+ (Chen et al., 2018) during training. Segmentation evaluation is reported on ACDC (Sakaridis et al., 2021) validation set.

Augmented images (%)	road IoU↑	sidewalk IoU↑	mIoU↑
100	36.43	35.26	39.83
66	60.43	43.81	46.84
50	68.78	45.79	**50.31**
20	75.85	44.60	47.36
10	**80.56**	**49.52**	47.64
0 (none)	74.95	39.52	45.31

Augmentation Percentage. We study also how the percentage of augmented images shown to DeepLabv3+ network at training impacts performances on semantic segmentation in snowy conditions.

As indicated in Tab. 5, even if we achieve the best performances with an augmentation probability of 50% (+5% mIoU w.r.t. no augmentation), we still use for evaluation in Sec. 4.2 the model obtained with 10% for its higher accuracy on road and sidewalk – crucial for autonomous navigation tasks.

5 CONCLUSION

In this work, we proposed a patch-based image-to-image translation model which relies on a GAN backbone trained on patches and an optional VAE to interpolate non-linearly between domains. Along with the definition of *local domains*, we introduced a dataset-based geometrical guidance strategy to ease the patches extraction process. Our few-shot method outperformed the literature on all tested metrics on several tasks (lane degradation, snow addition, deblurring), and its usability has been demonstrated on proxy tasks. In particular, our translation pipeline led to higher performances on lane detection in scenes with degraded or absent markings and on semantic segmentation in snowy conditions.

REFERENCES

Arar, M., Ginger, Y., Danon, D., Bermano, A. H., and Cohen-Or, D. (2020). Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *CVPR*.

Baek et al., K. (2020). Rethinking the truly unsupervised image-to-image translation. *arXiv*.

Bhattacharjee, D., Kim, S., Vizier, G., and Salzmann, M. (2020). Dunit: Detection-based unsupervised image-to-image translation. In *CVPR*.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*.

Bruls, T., Porav, H., Kunze, L., and Newman, P. (2019). Generating all the roads to rome: Road layout randomization for improved road marking segmentation. In *ITSC*.

Cao, J., Hou, L., Yang, M.-H., He, R., and Sun, Z. (2021). Remix: Towards image-to-image translation with limited data. In *CVPR*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

Cherian, A. and Sullivan, A. (2019). Sem-gan: Semantically-consistent image-to-image translation. In *WACV*.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*.

Cord, A. and Aubert, D. (2011). Towards rain detection through use of in-vehicle multipurpose cameras. In *IV*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Golestaneh, S. A. and Karam, L. J. (2017). Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *CVPR*.

Gong, R., Li, W., Chen, Y., and Gool, L. V. (2019). Dlow: Domain flow for adaptation and generalization. In *CVPR*.

Halder, S. S., Lalonde, J.-F., and de Charette, R. (2019). Physics-based rendering for improving robustness to rain. In *ICCV*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Huang, H., Li, Z., He, R., Sun, Z., and Tan, T. (2018a). Introvae: Introspective variational autoencoders for photographic image synthesis. In *NeurIPS*.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018b). Multimodal unsupervised image-to-image translation. In *ECCV*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Jeong, S., Kim, Y., Lee, E., and Sohn, K. (2021). Memory-guided unsupervised image-to-image translation. In *CVPR*.

Kim, J., Kim, M., Kang, H., and Lee, K. (2020). U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. (2017). Fader networks: Manipulating images by sliding attributes. In *NeurIPS*.

Lee, H., Seol, J., and Lee, S.-g. (2021). Contrastive learning for unsupervised image-to-image translation. *arXiv*.

Lee, H.-Y., Tseng, H.-Y., Mao, Q., Huang, J.-B., Lu, Y.-D., Singh, M., and Yang, M.-H. (2020). Drit++: Diverse image-to-image translation via disentangled representations. *IJCV*.

Li, P., Liang, X., Jia, D., and Xing, E. P. (2018). Semantic-aware grad-gan for virtual-to-real urban scene adaption. *BMVC*.

Lin, C.-T., Wu, Y.-Y., Hsu, P.-H., and Lai, S.-H. (2020). Multimodal structure-consistent image-to-image translation. In *AAAI*.

Lin, J., Xia, Y., Liu, S., Zhao, S., and Chen, Z. (2021a). Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. *Neurocomputing*.

Lin, Y., Wang, Y., Li, Y., Gao, Y., Wang, Z., and Khan, L. (2021b). Attention-based spatial guidance for image-to-image translation. In *WACV*.

Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. In *NeurIPS*.

Liu, M.-Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz, J. (2019). Few-shot unsupervised image-to-image translation. In *ICCV*.

Liu, Y., De Nadai, M., Cai, D., Li, H., Alameda-Pineda, X., Sebe, N., and Lepri, B. (2020). Describe what to change: A text-guided unsupervised image-to-image translation approach. In *ACM MM*.

Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., and Van Gool, L. (2019). Exemplar guided unsupervised image-to-image translation with semantic consistency. In *ICLR*.

Ma, S., Fu, J., Wen Chen, C., and Mei, T. (2018). Dagan: Instance-level image translation by deep attention generative adversarial networks. In *CVPR*.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. (2017). Least squares generative adversarial networks. In *ICCV*.

Mejjati, Y. A., Richardt, C., Tompkin, J., Cosker, D., and Kim, K. I. (2018). Unsupervised attention-guided image-to-image translation. In *NeurIPS*.

Mo, S., Cho, M., and Shin, J. (2019). Instagan: Instance-aware image-to-image translation. *ICLR*.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Pan, X., Shi, J., Luo, P., Wang, X., and Tang, X. (2017). Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI*.

Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In *ECCV*.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

Patashnik, O., Danon, D., Zhang, H., and Cohen-Or, D. (2021). Balagan: Cross-modal image translation between imbalanced domains. In *CVPR Workshops*.

Pizzati, F., Cerri, P., and de Charette, R. (2021a). Comogan: continuous model-guided image-to-image translation. In *CVPR*.

Pizzati, F., Cerri, P., and de Charette, R. (2021b). Guided disentanglement in generative networks. *arXiv*.

Ramirez, P. Z., Tonioni, A., and Di Stefano, L. (2018). Exploiting semantics in adversarial training for image-level domain adaptation. In *IPAS*.

Romera, E., Bergasa, L. M., Yang, K., Alvarez, J. M., and Barea, R. (2019). Bridging the day and night domain gap for semantic segmentation. In *IV*.

Saito, K., Saenko, K., and Liu, M.-Y. (2020). Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In *ECCV*.

Sakaridis, C., Dai, D., and Gool, L. V. (2021). Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*.

Schutera, M., Hussein, M., Abhau, J., Mikut, R., and Reischl, M. (2020). Night-to-day: Online image-to-image translation for object detection within autonomous driving by night. *IEEE T-IV*.

Shaham, T. R., Dekel, T., and Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. In *ICCV*.

Shen, Z., Huang, M., Shi, J., Xue, X., and Huang, T. S. (2019). Towards instance-level image-to-image translation. In *CVPR*.

Tang, H., Xu, D., Sebe, N., and Yan, Y. (2019a). Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*.

Tang, H., Xu, D., Yan, Y., Corso, J. J., Torr, P. H., and Sebe, N. (2019b). Multi-channel attention selection gans for guided image-to-image translation. In *CVPR*.

Tremblay, M., Halder, S. S., de Charette, R., and Lalonde, J.-F. (2020). Rain rendering for evaluating and improving robustness to bad weather. *IJCV*.

TuSimple (2021). Tusimple benchmark. In *https://github.com/TuSimple/tusimple-benchmark*.

Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., and Jawahar, C. V. (2019). Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*.

Wang, X., Yu, K., Dong, C., Tang, X., and Loy, C. C. (2019). Deep network interpolation for continuous imagery effect transition. In *CVPR*.

Wang, Y., Khan, S., Gonzalez-Garcia, A., Weijer, J. v. d., and Khan, F. S. (2020). Semi-supervised learning for few-shot image-to-image translation. In *CVPR*.

Wu, W., Cao, K., Li, C., Qian, C., and Loy, C. C. (2019). Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*.

Xu, M., Lee, J., Fuentes, A., Park, D. S., Yang, J., and Yoon, S. (2021). Instance-level image translation with a local discriminator. *IEEE Access*.

Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. (2019). Free-form image inpainting with gated convolution. In *ICCV*.

Yuan, Y., Chen, X., and Wang, J. (2020). Object-contextual representations for semantic segmentation. In *ECCV*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., and Jia, J. (2018). Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*.

Zheng, T., Fang, H., Zhang, Y., Tang, W., Yang, Z., Liu, H., and Cai, D. (2021). Resa: Recurrent feature-shift aggregator for lane detection. In *AAAI*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. (2017b). Toward multimodal image-to-image translation. In *NeurIPS*.

Zhu, P., Abdal, R., Qin, Y., and Wonka, P. (2020a). Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*.

Zhu, Z., Xu, Z., You, A., and Bai, X. (2020b). Semantically multi-modal image synthesis. In *CVPR*.

APPENDIX

In the experimental evaluation, we intentionally omit GAN metrics as they have important biases for two reasons we explain now. First, our method leverages only local domains translation while standard i2i applies a global transformation. Second, while we leverage high-level domain priors about local domains, we do not use any target images unlike standard i2i.

For completeness we still report GAN metrics against CycleGAN (Zhu et al., 2017a) and CycleGAN-15, which are trained on $\text{ACDC}_{\text{clear}} \mapsto \text{ACDC}_{\text{snow}}$ using respectively 400/400 or 400/15 source/target images. Comparatively, we *only* use the same 15 cherry picked $\text{ACDC}_{\text{snow}}$

Table 6: **Snow translation similarity on ACDC$_{\text{snow}}$.** GAN metrics on the snow addition task confirm the validity of our model. Without using any target images, our model yields acceptable results on FID, attaining even almost on-par performances with baselines on LPIPS.

Network	Training samples		FID↓	LPIPS↓
	clear	snow		
CycleGAN	400	400	**110.30**	**0.6225**
CycleGAN-15	400	15	111.29	0.6271
Ours	0	15	123.14	0.6283

images. The quantitative evaluation is obtained by performing roads and sidewalks translation of 100 ACDC$_{\text{clear}}$ images relying on segmentation masks from OCRNet (Yuan et al., 2020) pretrained on Cityscapes (Cordts et al., 2016). Since ACDC provides images (weakly) paired, we compute the pair-wise average LPIPS metric between each fake translation and its paired real snow image. We also evaluate FID between the fake and real snow datasets. It is important to note that we do not seek to outperform the baselines since they have access to ACDC$_{\text{clear}}$ images while our method does not. Results in Tab. 6 however show we perform reasonably good given the additional domain gap, even on par with baselines on LPIPS metric.