ManiFest: manifold deformation for few-shot image translation - supplementary material -

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

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

In this document, we provide additional insights about the training of ManiFest and details for the comparison with the baselines (Sec. 1). We also include more ablation studies (Sec. 2) about the role of anchors, the impact of GERM and more qualitative results on feature consistency. Finally, we study the deformed manifold resulting from our training (Sec. 3), we discuss limitations (Sec. 4), and we include more results on translation tasks (Sec. 5). For additional visualizations, please refer to the provided supplementary video.

1 Architectures and Training

ManiFest training Here we include details on ManiFest training procedure. Being GERM a residual correction, the network could use it to compensate an unrealistic suboptimal $\tilde{s}_w$ (*i.e.*, the image obtained from the style interpolation of WMI, see Sec. 3.2 and Eq. (2) of the main paper), which would consist of a naive pitfall in the ManiFest training. This will lead the network to encode strong corrections in $\tilde{s}_r$, with dramatic color changes. This depends mainly on training initialization. To avoid this behavior, we want $\tilde{s}_w$ to resemble as much as possible $\mathcal{T}$, so we additionally apply $\mathcal{L}_{\text{style}}$ also to $\tilde{s}_w$. In this way, residuals will be forced to encode only minor corrections to the discovered interpolated style, ultimately balancing the training procedure.

For the multi-target generator behavior, we follow StarGANV2 [1] and apply a multi-head mechanism to our style encoder, in the same way as in the mapping network.

We train ManiFest for 150,000 iterations, using Adam with $\beta_1 = 0.5, \beta_2 = 0.999$ with learning rate $1e-4$ for the network and $1e-2$ for w. For the patch-based discriminator, we use patches of size 64×64 and rotation up to 360°. We apply horizontal flipping as data augmentation.

Baselines training We now detail the adaptation of the FUNIT [3] and COCO-FUNIT [6] baselines. We decrease the number of downsampling blocks in both the content and style encoders to 2, use a single MLP block, and reduce the number of residual blocks to 4 following MUNIT [2]. We weight the reconstruction loss to 10. Please note that training with default hyperparameters leads to training divergence in our tasks. We use the official Imaginaire library for training[1].

[1] https://github.com/NVlabs/imaginaire, under NVIDIA License-NC

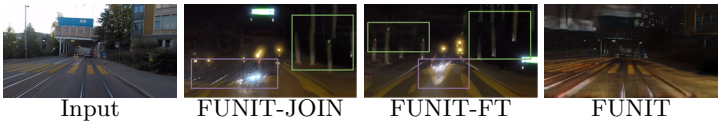


Fig. 1: We investigate different few-shot adaptation methods for FUNIT, based on finetuning on target (FUNIT-FT) or joint training with anchors (FUNIT-JOIN). Visual results for night (see fig. below) show that overfitting leads to loss of context (scene not recognizable) and artifacts (patterns on the street, poles), while FUNIT keeps scene structure.

FUNIT+LGFS	Day↦Night	Day↦Twilight	Clear↦Fog
(G)eneral	129.76 / 0.569	73.53 / 0.506	122.08 / 0.591
(E)xemplar	121.17 / 0.547	70.01 / 0.503	123.34 / 0.590

Table 1: FUNIT+LGFS results on the proposed translation tasks. While improving results, our few-shot learning strategy is best when coupled with WMI.

EGSC-IT has been trained with the official code and with the provided default configurations[2]. WCT^2 does not require retraining[3].

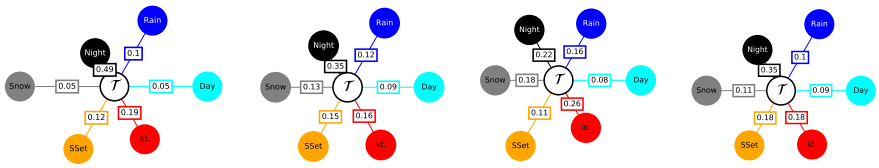
Baselines training with few-shot Different from ManiFest, FUNIT/COCO-FUNIT exploit few-shot *inference adaptation* as described in [3,6] but are not designed for few-shot *during training*. This is because they use a discriminator head for each domain, hence training on $\mathcal{T}$ will lead to overfitting. We nevertheless adapt FUNIT to investigate the best few-shot training adaptation setup: training on $\mathcal{A}_m$ and finetuning on $\mathcal{T}$ (FUNIT-FT) or training on joint $\{\mathcal{T}, \mathbb{A}\}$ (FUNIT-JOIN), and compare with the original setup (FUNIT). Effects of discriminator overfitting in Fig. 1 are evident and show that training with few-shot is less appropriate for FUNIT/COCO-FUNIT.

Baselines training with LGFS We adapt our best baseline FUNIT adding the LGFS loss on top of the original training procedure to understand the modularity of our approach, and coin this variant FUNIT+LGFS. We report FID/LPIPS in Tab. 1 following our standard evaluation pipeline. Results show improvements over FUNIT, but ManiFest (cf. main paper Tab. 1a,b,c) remains closer to target as it combines WMI and LGFS.

2 Ablation studies

Anchors To understand how estimating the Weighted Manifold Interpolation (WMI, main paper Sec. 3.2) weights $\tilde{w}$ is impacted by the different anchors, we

[2] https://github.com/charliememory/EGSC-IT
[3] https://github.com/clovaai/WCT2, under MIT license



(a) Day ↦ Night (b) Day ↦ Twilight (c) Clear ↦ Fog (d) Cross-dataset

Fig. 2: Illustration of the learned weights $\tilde{w}$ for each target domain of the multi-anchor *"All"* setup (see main paper, Sec. 4.5). The proximity of $\mathcal{T}$ to visually similar anchor domains proves the correct behavior of WMI. Distances depend on the weight of each domain reported in squares. "SSet" refers to the "Sunset" anchor.

Fig. 3: Increased feature consistency is evident if we compare LGFS-only (*c.f.* main paper, Sec. 4.5) to encoded target in the manifold spanned by anchors (Ours). In night rendering, the sky is better darkened, while the road gains more realistic illumination (e.g. cols 3, 5, 7).

analyze the weight assigned to each domain style representation in the multi-anchor (*"All"*) setup, as explained in main paper, Sec. 4.5.

Fig. 2 illustrates the distance maps between $\mathcal{T}$ and all anchor domains used for training. We observe that the anchor weights indeed correlate to the visual similarity with the target domain. For instance, the *Night* synthetic anchor is assigned very a high weight in the Day ↦ Night task, while its importance decreases in Clear ↦ Fog. This illustration demonstrates that the estimated weights $\tilde{w}$ make intuitive sense, and also opens new doors for understanding relationships between different domains in an unsupervised manner.

Feature consistency In Fig. 3, we provide additional qualitative results comparing ManiFest to the "LGFS-only" setup described in the main paper, see Sec. 4.5 and Fig. 7. This further illustrates the feature consistency obtained with the WMI in ManiFest, since all translations "LGFS-only" have unrealistic traits (e.g. the sky not uniformly translated, the road too dark, etc.). ManiFest, instead, exploits the manifold spanned by anchors to learn consistent features transformations, resulting in better translations.

GERM performances One could argue that the exemplar superior performances could be unrelated to the exemplar behavior, and could be attributed, for example,

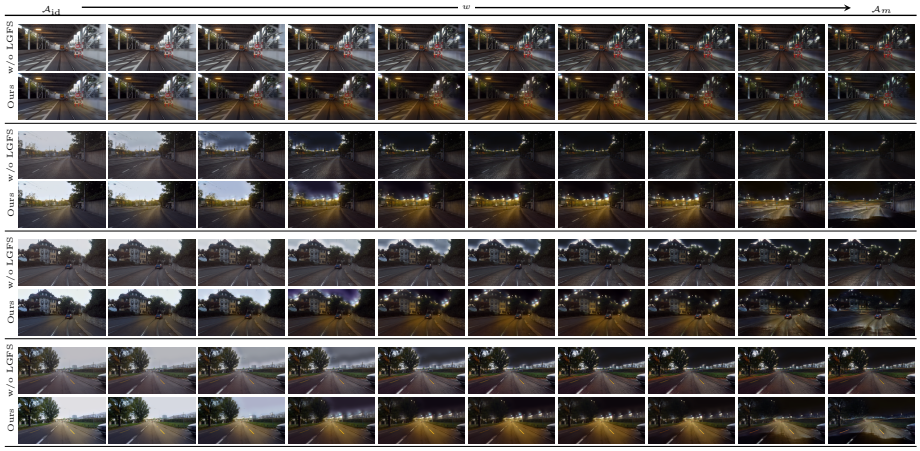


Fig. 4: Our manifold deformation strategy can deform the manifold spanned by anchors to inject $\mathcal{T}$ appearance in between anchor styles. The same network trained without LGFS, instead, only linearly interpolates between anchors.

to noise. To show that it is our exemplar style extraction path that is improving performance, we evaluate LPIPS on Day $\mapsto$ Night by selecting a random image as reference for LPIPS evaluation instead of the correctly-paired one (we refer to this as *exemplar-rand*). This experiment results in exemplar-rand / general / exemplar LPIPS 0.593 / 0.535 / **0.525**. Since LPIPS is evaluated exploiting paired reference images, removing the conditioning on those dramatically increases LPIPS, thus demonstrating that paired image style mimicking with exemplar style transfer is useful for GAN metrics. This also shows that the general style extraction is approximating well the whole few-shot set since it achieves good performance without explicit conditioning. We also evaluated the performances of our learned general style by comparing that with the one obtained averaging all style codes extracted in $\mathcal{T}$ (*avg-exemplar*). We obtained for avg-exemplar/general on Day $\mapsto$ Night FID 82.11/**81.01** and LPIPS 0.540/**0.535**.

Distance between domains If Anchor $\mathcal{A}$ and Target $\mathcal{T}$ were too close, we would just expand the few-shot training set. For this reason, we report FID of Source/Target ($\mathcal{S}/\mathcal{T}$), Anchor/Target ($\mathcal{A}/\mathcal{T}$) and Source/Anchor ($\mathcal{S}/\mathcal{A}$). $\mathcal{S}/\mathcal{T}$–$\mathcal{A}/\mathcal{T}$–$\mathcal{S}/\mathcal{A}$ is 150.70–141.62–161.26 for Day$\mapsto$Night, 104.80–105.16–123.65 for Day$\mapsto$Twilight, 121.84–179.36–152.97 for Clear$\mapsto$Fog. All distances being comparable, this shows that $\mathcal{A}$ is not expanding training data.

Cardinality of $\mathcal{A}$ We test small-scale anchor set on Day$\mapsto$Night varying $|\mathcal{A}|$ to understand how much we require large-scale anchor sets, and get FID/LPIPS **81.01/0.535** for $|\mathcal{A}| = 3090$, 82.46/0.536 for $|\mathcal{A}| = 500$, and 83.85/0.554 for $|\mathcal{A}| = 100$. Note that even smaller $|\mathcal{A}|$ **outperform baselines** (main paper,

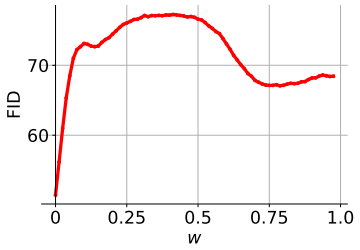


Fig. 5: We study the impact of our deformation strategy with respect to a manifold learned without $\mathcal{T}$ injection. Modifying w values comes with greater feature distance, demonstrating that $\mathcal{T}$ is injected between anchors and impacts less the GAN performances in rendering the anchor styles.

Fig. 6: Illustration of limitations on the ACDC Clear↦Rain and Clear↦Snow tasks. While correctly approximating $\mathcal{T}$, our output images lack specific traits such as reflections or snowy sidewalks.

Tab. 1a). 'LGFS-only' (main paper, Fig. 7) outperforms baselines **without** $\mathcal{A}$, but loses realism.

3 Manifold deformation

Fig. 4 provides visual results on the injection of $\mathcal{T}$ in between anchors. We compare ManiFest with a version of our multi-task network trained without LGFS, thus preventing the injection of $\mathcal{T}$ in the manifold spanned by $\mathbb{A}$. As visible, the "w/o LGFS" version only linearly interpolates between domains, while ManiFest injects $\mathcal{T}$ appearance in the manifold, while only slightly modifying the learned $\mathcal{A}_{\text{id}}$ and $\mathcal{A}_m$ appearance. This is also visible in the supplementary video.

To quantify deformation, we divide the learned manifold into 100 bins and measure the FID for each of them, comparing images of the "w/o LGFS" and "Ours" trainings for different values of w. Results are presented in Fig. 5, which shows that the deformation reaches a maximum between anchors, and decreases while approaching either $\mathcal{A}_{\text{id}}$ or $\mathcal{A}_m$.

4 Limitations

The main limitation of ManiFest is the need to retrain for adapting to different few-shot sets (full training takes around 1.5 days on a RTX 2080, with 1.11 sec/iter), but the entire pipeline is shown to generalize sufficiently to reduce

Fig. 7: ManiFest is adaptable to few-shot semantic transformations as Cat ↦ Fox, where the structure of the output image is changed (look for instance at the nose of generated foxes).

this need in the tested tasks. While this lowers applicability for entertainment purposes (*e.g.* image editing), training time is negligible for true rare few-shot scenarios (main paper, Sec. 4.4). Furthermore, others [5] require retraining for different $\mathcal{T}$. Another limitation is related to the local transformation capabilities of the network. We tested additional tasks as Clear↦Rain and Clear↦Snow using ACDC corresponding images (see Fig. 6). As visible, even if ManiFest correctly reproduces target images general appearance, it fails to render small but important traits for the scene realism such as reflections or snowy sidewalks. We hypothesize these features require additional contextual understanding to be rendered and propose to inject semantic guidance in ManiFest as future development.

5 Additional results

Semantic transformations We purposely designed ManiFest to be most effective on unstructured environments on which others fail. Nevertheless, in Fig. 7 we demonstrate below that it also works for semantic transformations on AFHQ-v2 [1] for the few-shot cat ↦ fox, using "dog" as anchor.

WCT^2 results on rare few-shot Due to space reason, we report in Fig. 8 WCT^2 results on the proposed Mountain ↦ Volcano and Day ↦ Aurora tasks proposed in the main paper, Sec. 4.4. As visible, WCT^2 creates more artifact and less faithful translations than ManiFest when applied in the exemplar scenario.

Segmentation performances on other domains We report additional segmentation downstream task performances following main paper, Fig. 5 setup. MUNIT/Ours mIoU is 32.81/**37.70** for Clear ↦ Fog, 37.22/**40.48** for Clear ↦ Rain, 37.67/**39.39** for Clear ↦ Snow (see limitations setups in Sec. 4). Day ↦ Twilight

(a) Mountain ↦ Volcano (b) Day ↦ Aurora

Fig. 8: We extend here Fig. 6 from main paper with WCT^2. Compared to ManiFest, WCT^2 generated less realistic mappings style mapping due to less evident semantic consistency. This is especially evident in the first column of a, where the mountain has unrealistic red regions.

lacks semantic labels so we show sample in Fig. 9. Although we list Rain and Snow as limitations in Sec. 4, segmentation still benefits from our better domain alignment compared to MUNIT [2].

Additional qualitative outputs We include additional qualitative results for the Day ↦ Night (Fig. 10), Clear ↦ Fog (Fig. 11) and Day ↦ Twilight (Fig. 12) tasks, along with additional segmentation results (Fig. 13) with the setups described in Sec. 4.3 of the main paper. All results are coherent with the ones described in the main document.

In the supplementary video, we demonstrate Day ↦ Night generalization to a YouTube driving sequence.

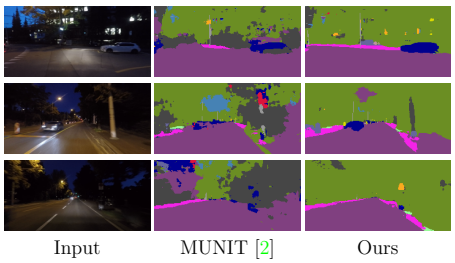


Fig. 9: Qualitative comparison for segmentation on Dark Zurich twilight sequence trained on translated images with MUNIT or ManiFest.

.

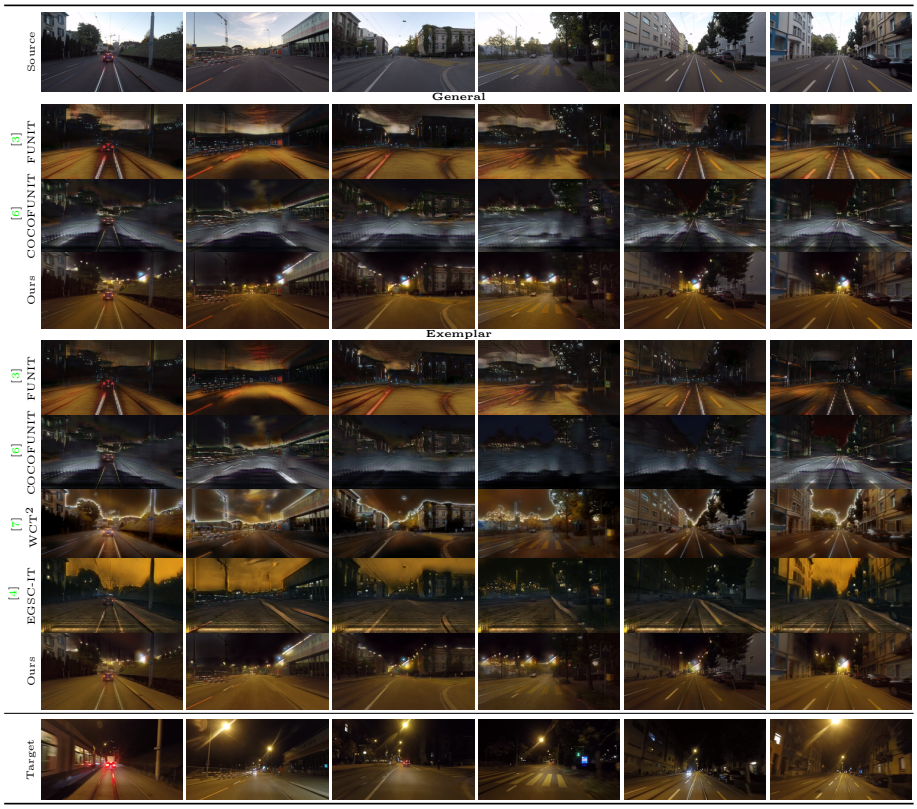


Fig. 10: Additional Day ↦ Night qualitative evaluation.

Fig. 11: Additional Clear ↦ Fog qualitative evaluation.

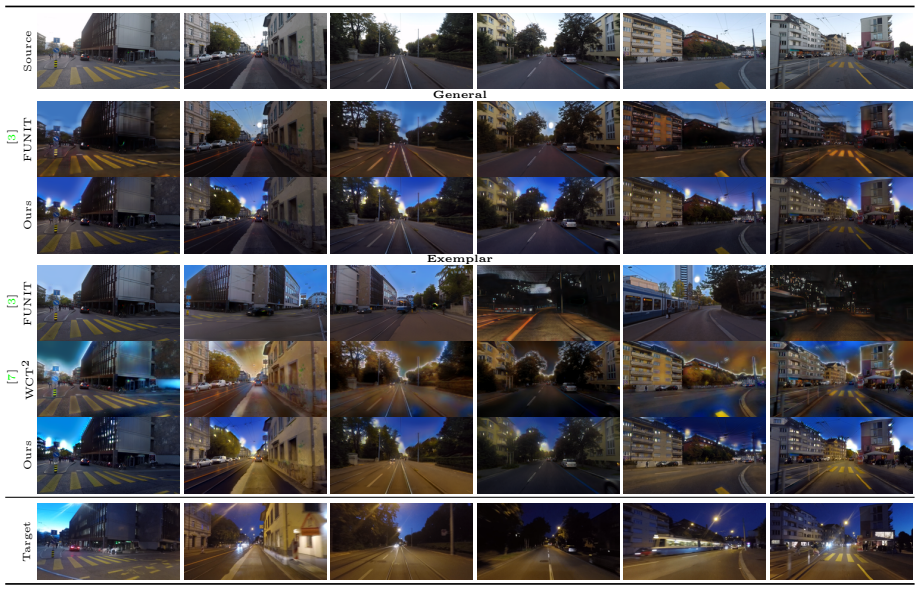


Fig. 12: Additional Day ↦ Twilight qualitative evaluation.

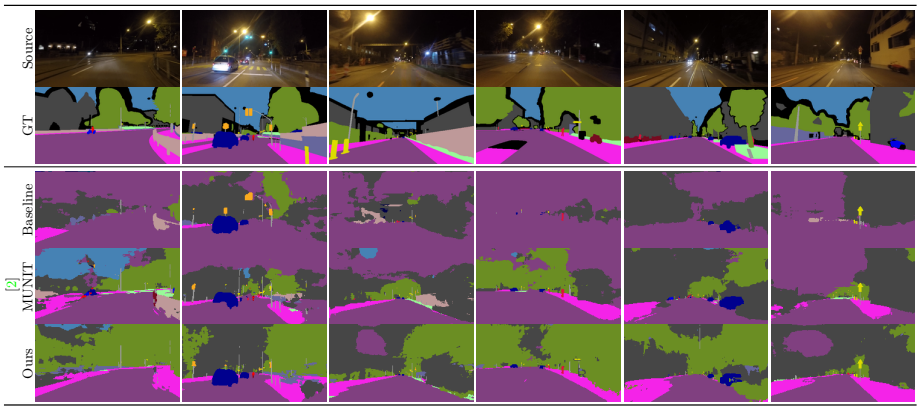


Fig. 13: Additional qualitative evaluation on semantic segmentation on ACDC night.

References

1. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR (2020) 1, 6
2. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018) 1, 6, 7, 8, 10
3. 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, 2, 7, 8, 9
4. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. In: ICLR (2019) 8
5. 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) 6
6. Saito, K., Saenko, K., Liu, M.Y.: COCO-FUNIT: Few-shot unsupervised image translation with a content conditioned style encoder. In: ECCV (2020) 1, 2, 8
7. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: ICCV (2019) 7, 8, 9