arXiv:2103.06879v2 [cs.CV] 8 Apr 2021

CoMoGAN: continuous model-guided image-to-image translation

Fabio Pizzati
Inria, Vislab
fabio.pizzati@inria.fr

Pietro Cerri
Vislab
pcerri@ambarella.com

Raoul de Charette
Inria
raoul.de-charette@inria.fr

Abstract

CoMoGAN is a continuous GAN relying on the unsupervised reorganization of the target data on a functional manifold. To that matter, we introduce a new Functional Instance Normalization layer and residual mechanism, which together disentangle image content from position on target manifold. We rely on naive physics-inspired models to guide the training while allowing private model/translations features. CoMoGAN can be used with any GAN backbone and allows new types of image translation, such as cyclic image translation like timelapse generation, or detached linear translation. On all datasets, it outperforms the literature. Our code is available in this page: https://github.com/cv-rits/CoMoGAN.

1. Introduction

Image-to-image (i2i) translation networks learn translations between domains, applying to the context of source images a target appearance learned from a dataset. This enables applications such as neural photo editing [75, 32, 21, 48, 6], along with robotics-oriented tasks as time-of-day or weather selection [74, 47, 46, 13, 61], domain adaptation [18, 40, 29, 60], or others. Despite impressive leaps forward with unpaired [75, 32], multi-target [9, 65], or continuous [64, 14] i2i, there are still important limitations. Specifically, to learn complex continuous translations existing works require supervision on intermediate domain points. Also, they assume piece-wise or entire linearity of the domain manifold. Such constraints can hardly meet cyclic translations (e.g. daytime) or continuous ones costly or impractical to label (e.g. fog, rain).

Instead, we introduce CoMoGAN, the first i2i framework learning non-linear continuous translations with unsupervised target data. It is trained using simple physics-inspired models for guidance, while relaxing model dependency via continuous disentanglement of domain features. An interesting resulting property is that CoMoGAN discovers the target data manifold ordering, unsupervised. For evaluation we propose new translation tasks, shown in Fig. 1, being either cyclic/linear, attached/detached from source. Our contributions are:

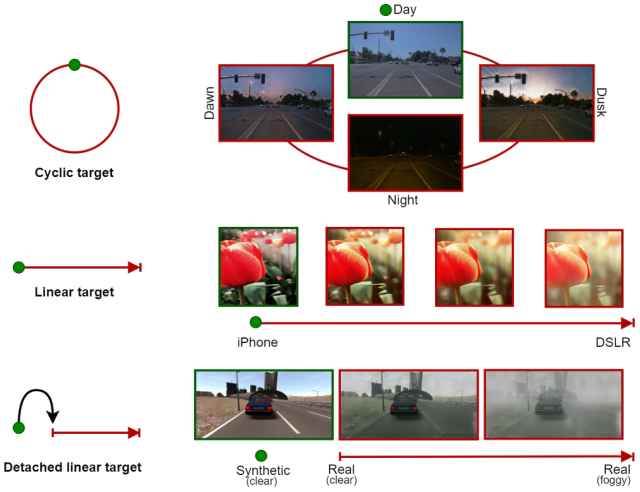


Figure 1: Detaching from traditional i2i translation, we are interested in *continuous* mapping from source domain (green point) to a target domain (red lines), in single- or multi- modal setup. A key feature of our proposal, is unsupervised reorganization of the data along a functional manifold (top: cyclic, middle/bottom: linear). We leverage lighting translations from day images (top), shallower depth of field from in-focus images (middle), or synthetic clear images to realistic foggy images (bottom).

- a novel model-guided setting for continuous i2i,
- CoMoGAN: an unsupervised framework for disentanglement of continuously evolving features in generated images, using simple model guidance,
- a novel Functional Instance Normalization (FIN) layer,
- the evaluation of CoMoGAN against recent baselines and new tasks, outperforming the literature on all.

2. Related works

Differently from early i2i [22], the seminal work in [75, 70] enabled unpaired source/target training. Building on it, multi-modal [21, 76] or multi-target [8, 9, 65, 2] i2i appeared. Performance was also boosted with additional supervision [55, 5, 39, 27, 58, 7, 78, 77, 30, 36, 45, 41, 35].

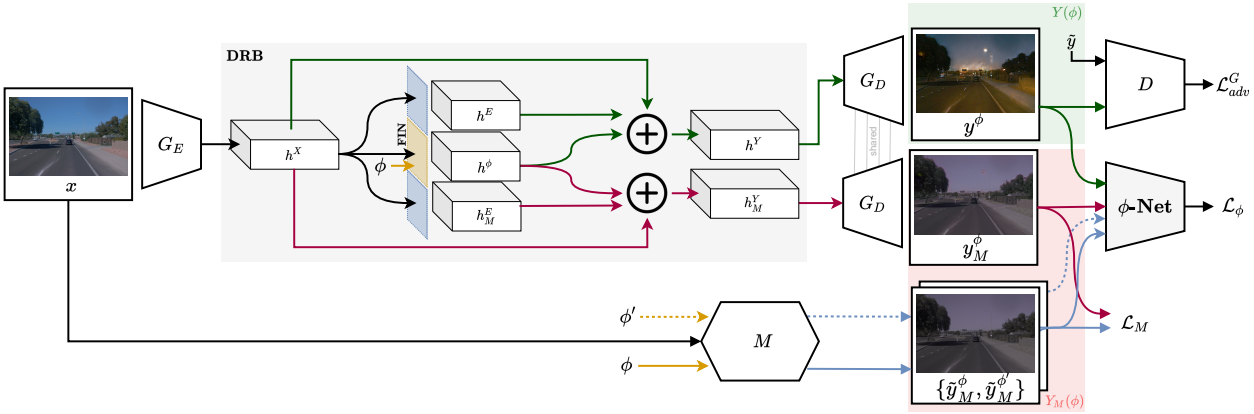


Figure 2: CoMoGAN enables unsupervised continuous translation, being end-to-end trainable, and architecture agnostic. Our Disentanglement Residual Block (DRB) – placed between encoder/decoder (G_E/G_D) – uses new Functional Instance Normalization (FIN, yellow layer) to learn manifold reshaping and continuous translation, guided with simple physics-inspired model M. For losses ($\mathcal{L}$), on top of standard ones we optimize model reconstruction ($\mathcal{L}_M$) and manifold consistency ($\mathcal{L}_\phi$) by enforcing manifold distances between GAN output and model outputs $\{\phi, \phi'\}$ with a pair-wise estimator (ϕ-Net).

Model-guided translation. Models can be exploited to improve i2i. In [61], they hybrid a physics based rendering [15] with GANs to enable controllable rainy translation. Similarly, [46] disentangles occlusions by injecting models at training. All these rely on model integration, rather than guidance. Models could influence many training aspects, in the form of output space conditions [49], loss functions [25] or ad-hoc data augmentation [68]. They have been used extensively for image restoration [43, 28, 69], but rarely for GAN image synthesis. Still, [23] uses simple models to learn basic image transformation (rotation, brightness, etc.).

Disentangled representations. Disentanglement is commonly used to gain control on generation by separating image content and style [21, 26, 24, 44]. Others aim at controlling output images granularity [56] or specific features, as blur [34] or view-points [42]. Some exploit disentanglement for few-shot generalization capabilities [33, 52]. Domain features disentanglement also unifies representations across domains [66, 31]. While some do not use labels at all [3, 4], none of them learn translation sequentiality.

Continuous image translation. A common practice for continuous i2i is to use intermediate domains by weighting discriminator [14, 13], using losses for middle states [65], or mixing disentangled styles representations [9, 50]. Attribute vectors interpolation [67, 71, 37] enables continuous control of several features. Others continuously navigate latent spaces with discovered paths [6, 12, 23]. Finally, feature [63] or kernel [64] interpolation were proposed. Still, they assume linear interpolation – not always valid (e.g. day to night include dusk). GANimation [48] instead, use non-linear interpolations but require intermediate domain labels.

3. CoMoGAN

Instead of a point-to-point mapping ($X \mapsto Y$), CoMoGAN learns a continuous domain translation controlled by ϕ, that is $X \mapsto Y(\phi)$. Training uses source data (at fixed ϕ_0) and unsupervised target data (unknown ϕ). It reshapes the data manifold guided by naive physics-inspired models (e.g. tone-mapping, blurring, etc.). Rather than mimicry, we relax the model and let the networks discover private image features via our disentanglement of output, ϕ, and style.

Fig. 2 is an overview of our architecture-agnostic proposal. It relies on three key components. We first introduce Functional Instance Normalization layer (Sec. 3.1) which enables ϕ-manifold reshaping. Second, our Disentanglement Residual Block (Sec. 3.2) in charge of ϕ disentanglement in input data. Finally, we detail ϕ-Net, a pair-wise ϕ regression network (Sec. 3.3) which enforces manifold distances consistency.

Model guidance. We guide the learning with *simple* non-neural models $M(x, \phi)$, x the source image. Thus, following the intuition that target manifold can be discovered with coarse guidance: night resembles *dark* day, fog looks like a *blurry gray* clear image, etc. We depart from the need of complex physical guidance since we disentangle shared and private features from model/translation which enables discovering complex non-modeled features (e.g. light sources at night). Models are described in Sec. 4.1 and supp.

3.1. Functional Instance Normalization (FIN)

To take advantage of our model guidance which is continuous by nature, we must allow our network to encode

ϕ continuity. To do so, we build on prior Instance Normalization (IN) which allows carrying style-related information [62, 20]. It writes for input x,

$$\text{IN}(x) = \frac{x-\mu}{\sigma}\gamma + \beta, \tag{1}$$

where μ and σ are input feature statistics, and γ and β learned parameters of an affine transformation. As an extension, we propose Functional Instance Normalization (FIN)

$$\text{FIN}(x,\phi) = \frac{x-\mu}{\sigma}f_\gamma(\phi) + f_\beta(\phi), \tag{2}$$

where instead of learning a unique value of affine transformation parameters, we learn the distribution of transformations f_γ and f_β. The intuition is to shape the ϕ-manifold based on how the transformation evolves. Compared to others [14], this allows us to interpret better the learned manifold. Depending on the nature of $Y(\phi)$, we can encode FIN layer accordingly. In this work, we investigate linear and cyclic encoding. Linear encoding is commonly encountered, and assumes reorganizing features linearly. For instance, considering adverse weather phenomena, severe conditions (e.g. thick fog) are always positioned after light ones (i.e. lite fog). We model linear FIN parameters as

$$\begin{aligned} f_\gamma(\phi) &= a_\gamma\phi + b_\gamma, \\ f_\beta\phi &= a_\beta\phi + b_\beta, \end{aligned} \tag{3}$$

with $\{a_\gamma, a_\beta, b_\gamma, b_\beta\}$ the learnable parameters of the layer.

Conversely, some translations path loop back to source, as it happens with daylight, which is *cyclic* by nature going from Day to Dusk $\mapsto$ Night $\mapsto$ Dawn and Day again. In this case, we encode cyclic FIN layer with parameters

$$\begin{aligned} f_\gamma(\phi) &= a_\gamma cos(\phi) + b_\gamma, \\ f_\beta(\phi) &= a_\beta sin(\phi) + b_\beta. \end{aligned} \tag{4}$$

3.2. Disentanglement Residual Block (DRB)

The pitfall of strict model-dependency is that the GAN will only learn to mimic the model. To prevent that, we must allow target domain $Y(\phi)$ and *model* domain $Y_M(\phi)$ to have shared *modeled* features Y^ϕ but also private *non-modeled* features Y^E and Y^E_M, respectively. This writes

$$\begin{aligned} Y(\phi) &= \{Y^\phi, Y^E\}, \\ Y_M(\phi) &= \{Y^\phi, Y^E_M\}. \end{aligned} \tag{5}$$

We enable private features in either domain with our Disentanglement Residual Block (DRB, shown in Fig. 2) whose goal is to extract disentangled representations for a given ϕ. The DRB is composed of residual blocks mapping the encoder feature map h^X to the disentangled representations of output images. Let $y^\phi \in Y(\phi), y^\phi_M \in Y_M(\phi)$, we have

$$\begin{aligned} \text{DRB}(h^X,\phi) &= \{h^Y, h^Y_M\}, \\ y^\phi = G_D(h^Y), &\quad y^\phi_M = G_D(h^Y_M). \end{aligned} \tag{6}$$

The DRB works as follows. Following Fig. 2, the input representation h^X is processed by residual blocks, each one extracting features associated with the atomic ones previously introduced, such as $Y^\phi, Y^E, Y^E_M \longleftrightarrow h^\phi, h^E, h^E_M$, one per residual. In particular, the residual block for h^ϕ extraction uses our FIN layers for normalization to encode continuous features. The hidden latent representations h^Y and h^Y_M are obtained from summation of the disentangled features and h^X to ease gradient propagation as in [16]. In formulas,

$$\begin{aligned} h^Y &= h^\phi + h^E + h^X, \\ h^Y_M &= h^\phi + h^E_M + h^X. \end{aligned} \tag{7}$$

Intuitively, for optimization we need feedback from both real data similarity and mimicking of the model output. While the first must rely on adversarial training due to the use of unpaired images, we can enforce reconstruction on the paired modeled $\tilde{y}^\phi_M = M(x,\phi)$. Assuming LSGAN [38] training and discriminator D, we obtain

$$\begin{aligned} \mathcal{L}^G_{adv} &= ||D(y^\phi) - 1||_2, \\ \mathcal{L}_M &= ||y^\phi_M - \tilde{y}^\phi_M||_1. \end{aligned} \tag{8}$$

Minimization of $\mathcal{L}^G_{adv}$ and $\mathcal{L}_M$ during the generator update step enables disentanglement of h^E and h^E_M.

3.3. Pairwise regression network (ϕ-Net)

The DRB enforces both disentanglement and manifold shape at a feature level, but it requires ad-hoc training strategies to actually disentangle also continuous features for real images and not fall into easy pitfalls, e.g. the network only exploiting h^E for target translation ignoring h^ϕ. Hence, we introduce a training strategy based on similarities which forces the network to both exploit extracted continuous information and follow the model guidance. Suppose an input image x, mapped to $x \mapsto y^\phi$ by the network. As shown in Fig. 2, we randomly sample ϕ and ϕ' and apply $M(.)$ to x, obtaining the couple $\{\tilde{y}^\phi_M, \tilde{y}^{\phi'}_M\}$. We use a CNN ($\phi$-Net) for domain similarity discovery. It takes as input a pair of images and regresses their ϕ differences, such as

$$\phi\text{-Net}(y^\phi, y^{\phi'}) = \phi - \phi' = \Delta\phi\,. \tag{9}$$

We jointly optimize ϕ-Net and generator (G) parameters in an end-to-end setting by enforcing consistency between real and modeled target domain images. In formulas,

$$\begin{aligned} \mathcal{L}^G_\phi &= ||\phi\text{-Net}(y^\phi, \tilde{y}^\phi_M)||_2 + ||\phi\text{-Net}(y^\phi, \tilde{y}^{\phi'}_M) - \Delta\phi||_2, \\ \mathcal{L}_{gt} &= ||\phi\text{-Net}(\tilde{y}^\phi_M, \tilde{y}^{\phi'}_M) - \Delta\phi||_2, \\ \mathcal{L}_\phi &= \mathcal{L}^G_\phi + \mathcal{L}_{gt}. \end{aligned} \tag{10}$$

$\mathcal{L}^G_\phi$ forces G to organize the manifold following the feedback of the physical model, ultimately resulting in generated y^ϕ and $\tilde{y}^\phi_M$ to be mapped to the same ϕ on the manifold

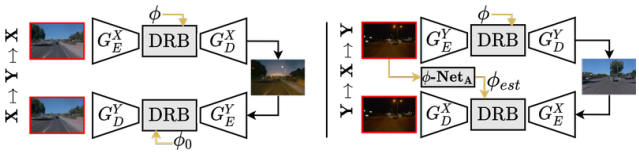


Figure 3: We enforce cycle consistency by injecting the source ϕ_0 in the $X \mapsto Y \mapsto X$ translation when reconstructing the original image. Also, for $Y \mapsto X \mapsto Y$ we position the input image at ϕ_{est} on the domain using our ϕ-Net$_A$ CNN trained unsupervised for ϕ regression.

discovered by ϕ-Net. That way, the network can identify that images follow some similarity criteria despite differences between model output and learned translation, leading to an organization of the latent space guided by the physical model. $\mathcal{L}_{gt}$ instead exploits modeled data only and thus is used to avoid training collapse. For linear FIN, we train on ϕ and $\Delta\phi$, though for cyclic one stability is increased by evaluating each loss on sin/cos projection of ϕ.

3.4. Training strategy

CoMoGAN is end-to-end trainable and can be used with any i2i framework by simply adding the DRB between encoder and decoder, with our losses. The final objective for the generator depends if source and target are detached, i.e. $X \not\subset Y$ (see Fig. 1 for visualization). If detached, the generator update step writes

$$\mathcal{L}^G = \mathcal{L}^G_{adv} + \mathcal{L}_M + \mathcal{L}_\phi \,. \tag{11}$$

For attached source/target, we enforce source (ϕ_0) identity:

$$\mathcal{L}^G = \mathcal{L}^G_{adv} + \mathcal{L}_M + \mathcal{L}_\phi + ||G(x, \phi_0) - x||_1 \,. \tag{12}$$

Either $\mathcal{L}^G$ definition is used, sometimes in conjunction with a regularization pairwise loss to ease training (cf. supp). Using real data ($\tilde{y}$) from target the discriminator minimizes

$$\mathcal{L}^D = \mathcal{L}^D_{adv} = ||D(y^\phi)||_2 + ||D(\tilde{y}) - 1||_2 \,.$$

Cycle consistency. In addition to $X \mapsto Y$, many networks perform $Y \mapsto X$ to preserve context with cycle consistency. To handle the latter, we insert a *shared* DRB between each encoder/decoder couple to benefit from multiple sources. This is illustrated in Fig. 3. We also use another unsupervised network, called ϕ-Net$_A$, that regresses ϕ on the target dataset. From above figure (left), because ϕ is injected in $X \mapsto Y$ transformation, we enforce a correct spreading of all ϕ values by adding $\mathcal{L}_{reg}$ to the generator objective, $\mathcal{L}_{reg} = ||\phi\text{-Net}_A(y^\phi) - \phi||_2$.

4. Experiments

We show the efficiency of CoMoGAN on new continuous image-to-image translation tasks $X \mapsto Y(\phi)$, where we consider source data to lie on a fixed point (ϕ_0) of the ϕ-manifold and *unknown* ϕ target data. The underlying optimization challenge is to learn simultaneously the ϕ-manifold and continuous image translation. Because continuous model-guided translation is new, we first describe our three novel translations tasks (Sec. 4.1) obtained by leveraging recent datasets [57, 51, 10, 15, 75]. Each task encompasses challenges of its own such as linear/cyclic target manifold, attached/detached manifolds (i.e. $X \subset Y$ or $X \not\subset Y$) and uni-/multi- modality. Specifically, we train with backbone MUNIT [21] (multi-modal) or CycleGAN [75] (uni-modal) and coin our alternatives CoMo-MUNIT and CoMo-CycleGAN, respectively. We evaluate the manifold organization (Sec. 4.2) and the translation quality (Sec. 4.3) from GAN metrics and proxy tasks. Continuous translation (Sec. 4.4) is evaluated separately and we conclude with ablation studies (Sec. 4.5). We mostly train with default backbone hyperparameters, more details are in supplementary.

4.1. Translation tasks

Day $\mapsto$ Timelapse. Using recent Waymo Open dataset [57], we frame the complex task of day to any time, thus learning *timelapse* passing through day/dusk/night/dawn. Waymo image labels are *only* used to split clear images into *source* {Day} and *target* {Dusk/Dawn, Night}, respectively obtaining train/val sets of 105307 / 28165 and 27272 / 7682 images. We train CoMo-MUNIT for multi-modality. To respect the cyclic nature of time we exploit cyclic FIN (Eq. 4) encoding $\phi \in [0, 2\pi]$, which maps to a sun elevation $\in [+30°, -40°]$. *For evaluation only*, we obtain ground truth elevation from astronomical models [1] with image GPS position and timestamp. For guidance, we exploit a simple day-to-night tone mapping [59] (Ω) interpolating with Hosek radiance model [19] (HSK) to account for gradual loss of color, and adding asymmetrical hue correction (corr) to account for temperature changes – i.e. at analog sun elevation dusk appears red-ish and dawn purple-ish –. The complete model is in the supplementary. It writes

$$M(x, \phi) = (1 - \alpha)x + \alpha\Omega(x, \text{HSK}(\phi) + \text{corr}(\phi)) + \text{corr}(\phi). \tag{13}$$

iPhone $\mapsto$ DSLR. We inspire from CycleGAN [75] by adapting their initial task to a continuous setup, learning the mapping of iPhone images with large depth of field to DSLR images with shallow depth of field. We also use the iphone2dslr flowers dataset [75], split in *source* 1182/569 and *target* 3325/480. We train this task with CoMo-CycleGAN for comparison, and use linear FIN (Eq. 3) where $\phi \in [0, 1]$ encodes the progression.

For guidance, we naively render blur by convolving ($*$) a Gaussian (G) which kernel size maps to ϕ. That is

$$M(x, \phi) = G(\phi) * x \,. \tag{14}$$

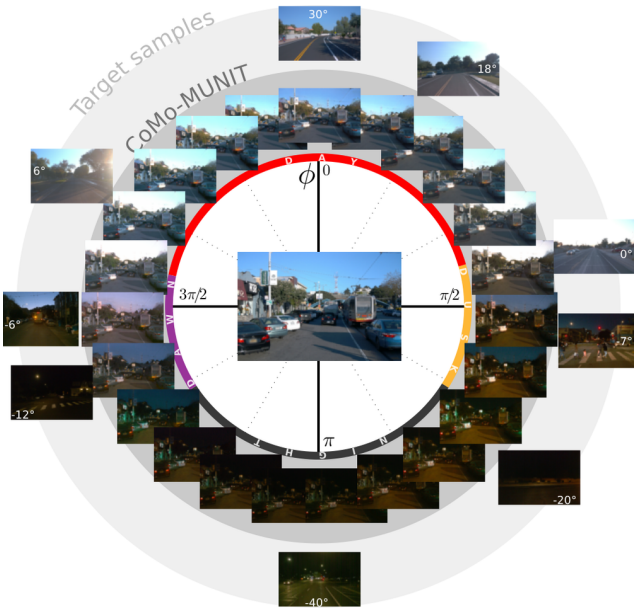


Figure 4: Translations (dark circle) of a source day image (center) exhibit both high variability and similarities with target data (outer circle) for which we report ground truth elevations. CoMo-MUNIT learned non-modeled visual features like frontal sun scenes resembling real ones (as in $\{0°, 6°, 18°\}$). Note that it discovered dawn/dusk and the stationary appearance of night, proving manifold quality.

Synthetic$_{\text{clear}}$ $\mapsto$ Real$_{\text{clear, foggy}}$. Here, we propose a detached source/target task, where we learn clear to foggy except that source is *synthetic* and target is *real* data. For *source*, we leverage spring sequences of synthetic Synthia dataset [51], split in 3497/959 images. As *target* we mix original Cityscapes [10] and 4 augmented foggy Weather Cityscapes [15] with max visibility distances {750m, 350m, 150m, 75m}. In target, each of the 5 Cityscapes version has 2975/500 images. We train here a CoMo-MUNIT with linear FIN layer (Eq. 3) and encode maximum visibility as $\phi \in [0, 1]$, i.e. visibility $\in [\infty, 70m]$. For guidance, we simply exploit the fog model of [15]. For the sake of space, models details, sample outputs and model experiments are provided in the supplementary.

4.2. Manifold organization

We evaluate the quality of the unsupervised manifold discovery using CoMo-MUNIT on the Day $\mapsto$ Timelapse. Fig. 4 shows a source day image (center) and our timelapse translations for uniformly sampled ϕ (middle circle). Apart from the appealing translations appearance, notice the network discovered important features like frontal sun (when the sun is close to the horizon), sunset/sunrise, material reflectance (at night), and the stable nighttime appearance. All these features are not in model $M(.)$ though present

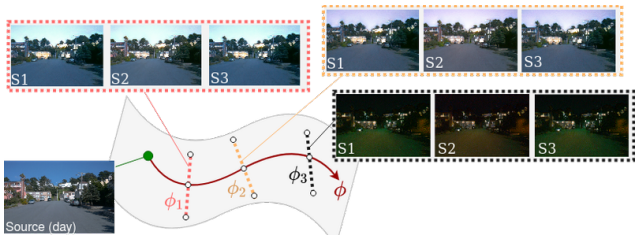


Figure 5: Translations along dimensions ϕ (red) and style (dotted). For a given ϕ, the styles vary slightly (notice hue and brightness), proving disentanglement of ϕ and style.

in target images (outer circle). This advocates the network disentangled model features and translation features. Note also that the top translation in Fig. 4 accurately resembles source, assessing that target is attached to source.

Quantitatively, we measure the manifold precision by regressing ϕ with our ϕ-Net$_A$ CNN (cf. Sec. 3.4) on real Waymo validation set, and compute the error w.r.t. ground truth elevations. We get a mean error of $19.8°$ (std $8.56°$) when *unsupervised* and $4.05°$ (std $4.20°$) if *supervised*. Even unsupervised, our manifold discovery is acceptable, and opens ways for unsupervised translations where ϕ ground truth would be impractical (e.g. rain, snow).

Disentangled dimensions. Because MUNIT is multimodal by design, it is important to assess CoMo-MUNIT properly disentangles ϕ from the style dimension of MUNIT. We do this by sampling ϕ and style. From Fig. 5, the latter evolve correctly on different axes, which was expected since ϕ is regulated by model-guided features. Again, using ϕ-Net$_A$, we regress ϕ values for 100 fixed ϕ translations each with 100 different styles, obtaining $1.06°$ ϕ-variance along the style dimension. This proves the orthogonality of ϕ and style manifolds.

4.3. Translation quality

GAN metrics. We measure the quality and variability of all translations task w.r.t. MUNIT and CycleGAN backbones, showcasing in Tab. 1 that we always perform better or on par. In the table, IS [54] evaluates image quality and diversity over all the dataset, CIS [21] over multimodal translations, and LPIPS [72] evaluates absolute diversity only. We conjecture our performance results of the higher degree of control we have, since we control ϕ features in a disentangled manner (i.e. extremely increasing variability), while entangled backbones lean towards the easiest translations. The InceptionV3 networks used for IS/CIS evaluation are trained on the source/target classification task. IS is evaluated on all validation set, while for CIS/LPIPS we follow [21] evaluation routine.

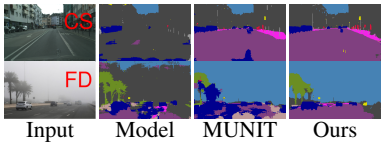
Task	Network	IS↑	CIS↑	LPIPS↑
Day ↦ Timelapse	MUNIT [21]	1.43	1.41	**0.583**
	CoMo-MUNIT	**1.59**	**1.51**	0.580
$\text{Syn.}_{\text{clear}}$ ↦ $\text{Real}_{\text{clear, foggy}}$	MUNIT [21]	**1.30**	1.02	0.493
	CoMo-MUNIT	**1.30**	**1.05**	**0.515**
iPhone ↦ DSLR	CycleGAN [75]	1.39	n.a.*	0.658
	CoMo-CycleGAN	**1.44**	**1.18**	**0.680**

* CIS is only applicable to multi-modal network.

Table 1: GAN metrics proves the benefit of our controllable ϕ generation, leading to on par or better quality/variability.

Translations	CS	FD	Mean
none (source)	10.9	10.1	10.5
Model [15]	19.9	21.5	20.7
MUNIT [21]	38.3	21.8	30.0
CoMo-MUNIT	**43.0**	**23.4**	**33.2**

(a) mIoU metric



Input Model MUNIT Ours

(b) Samples

Figure 6: Semantic segmentation on clear Cityscapes (CS) [10] and Foggy Driving (FD) [53] with PSPNet-50 [73] trained on clear Synthia (*source*), foggy physics *Model*, and $\text{Synthetic}_{\text{clear}}$ ↦ $\text{Real}_{\text{clear, foggy}}$ of MUNIT or CoMo-MUNIT. Noticeably, we outperform all on both clear (CS) and foggy (FD) dataset.

Semantic segmentation. We measure the effectiveness of our $\text{Synthetic}_{\text{clear}}$ ↦ $\text{Real}_{\text{clear, foggy}}$ translations in Fig. 6 by training PSPNet-50 [73] with either MUNIT or CoMo-MUNIT outputs. For comparison, we also train segmentation with clear *source* Synthia or physics-based foggy *model* [15] as for guidance. For MUNIT and CoMo-MUNIT, we employ a multi-modal style-sampling strategy [47] with 5 fixed styles. Additionally, for CoMo-MUNIT and *model* translations that allow it, we sample uniform ϕ. We follow [73] settings and train 150 epochs, using 3498 train images for each setup.

Tab. 6a reports the standard mIoU on shared Synthia-Cityscapes classes on real images from the validation set of Cityscapes [10] (CS, 500 images) and Foggy Driving [53] (FD, 101 images). While the transformation is subtle, it still reduces the domain shift, since even if *Model* significantly outperforms *source* but we beat all by additional margin of +4.7/+1.6/+3.2. Noticeably, we improve both on clear (CS) and foggy (FD) datasets showing CoMo-MUNIT preserved accurate clear *and* foggy translations. We speculate instead that MUNIT focuses on target dataset fog intensities which are discrete and may differ from FD, while our FIN layer enables continuous representation leading to better generalization. Qualitative evaluation on both datasets in Fig. 6b respects mIoU performances.

4.4. Continuous translation quality

To evaluate the continuity of the translations, we show uniformly spaced ϕ translations for Day ↦ Timelapse (Fig. 7, bottom row), Synthetic ↦ Real (Fig. 8) and iPhone ↦ DSLR (Fig. 9). For all, regardless of the backbone and task, our translations look appealing with our network discovering unique visual features *not* present in the model guidance. This is quite noticeable in DSLR (Fig. 9) which learned depth of field despite simple blurring guidance, or in the detached foggy experiment (Fig. 8) since translations encompass the desired real appearance with increasing fog.

4.4.1 Benchmark evaluation

We evaluate the challenging Day ↦ Timelapse with the literature. This is not trivial since our proposal is to the best of our knowledge the first continuous cyclic GAN. While *some* previous works could be adapted to cyclic translation (e.g. DLOW [14]) they *all* require intermediate labeled target points. Hence, to achieve a fair comparison compensating data scarcity in Waymo Open, we formulate timelapse as linear {Day, Dusk/Dawn, Night} for all baselines and randomly sample between Dusk or Dawn branch with our cyclic network. Please bear in mind that **all baselines are more supervised than ours** since they use intermediate Dusk/Dawn point while CoMoGAN discovers the manifold from unsupervised target data. We now detail the baselines.
StarGAN v2 [9] is a state-of-the-art multi-target i2i architecture learning multiple mapping from the same source point. We train it with official implementation on Day ↦ Dawn/Dusk ↦ Night path and use its style code disentanglement capability to enable continuous i2i.
DLOW [14] is continuous by design. We train it with 2 unimodal DLOW Day ↦ Dawn/Dusk and Dawn/Dusk ↦ Night. Note that it can be multi-target, but we already compare with the more recent StarGAN v2.
DNI [64] applies Deep Network Interpolation to interpolate among kernels of finetuned networks for continuous i2i. We adapt 2 baselines DNI-CycleGAN and DNI-MUNIT both trained on Day ↦ Dawn/Dusk ↦ Night.

Comparison. From Fig. 7, baselines (rows 1-4) either exhibit limited variability in interpolated points (StarGAN v2 / DNI) or unrealistic results (e.g. DLOW at night). A key limitation is that they rely on (piece-wise) linear interpolation preventing them from discovering the stationary aspect of night (last 3 cols). Conversely, CoMo-MUNIT (bottom row) translations are both realistic and stationary at night.

We also study the realism of all translations using the Frechet Inception Distance [17] (FID) to measure features distances between generated images and real ones. For that, we uniformly split the elevations range $[+30°, -40°]$ in 70 overlapping bins of 7° width, and compute each bin FID by comparing 100 translations and ad-hoc real images. We refer to this as "rolling FID", plotted in Fig. 10a. From the

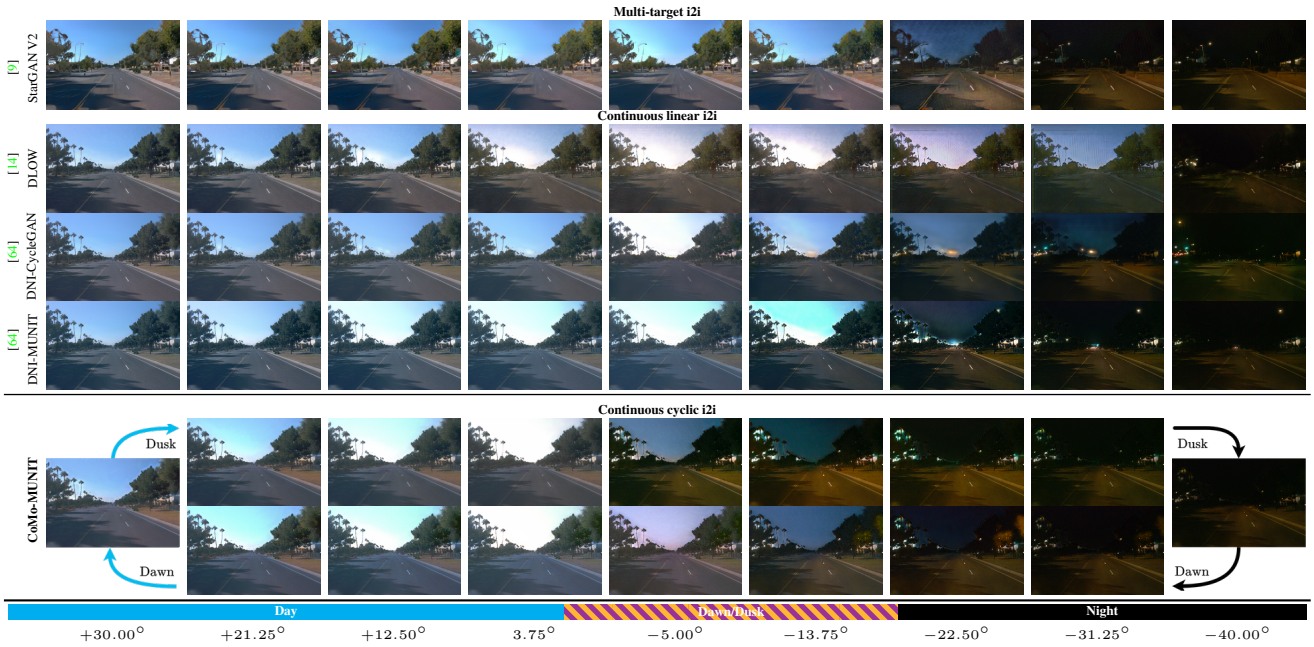


Figure 7: Day ↦ Timelapse translations. Baselines output unrealistic translations (e.g. DLOW [14]) or images with limited variability (StarGAN V2 [9]). DNI [64] is the best baseline, though our CoMo-MUNIT (last row) is the only cyclic one, outputs more variable images (e.g. at dusk/dawn) and discovered stable night *with less supervision*.

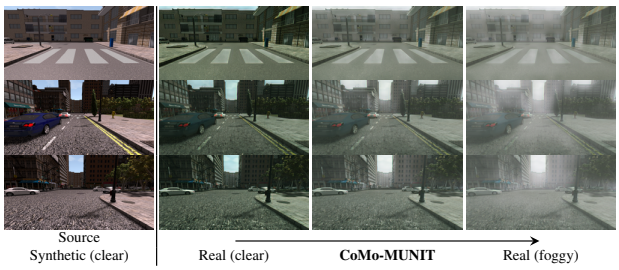


Figure 8: Sample $\text{Synthetic}_{\text{clear}} \mapsto \text{Real}_{\text{clear, foggy}}$ translations with CoMo-MUNIT. Note the complex detached source (Synthia [51]) and target (clear/foggy Cityscapes[10, 15]) setting. Still, clear translations correctly encompass Cityscapes stylistic appearance (notice texture and color).

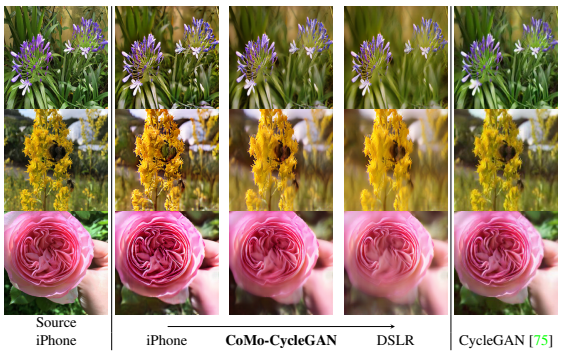


Figure 9: CoMo-CycleGAN translations on the iPhone ↦ DSLR task, using iphone2dslr dataset [75]. Despite naive blur guidance (Eq. 14), it learns continuous DSLR depth of field, while [75] outputs only target translations.

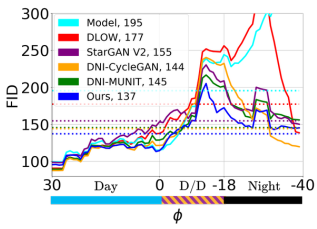
latter, our method outperforms others especially in complex intermediate conditions. Note the baselines performance at precise "dawn/dusk" center (where they are supervised) and how their FID degrade as they depart toward night (approx. $-18°$). Even if *unsupervised*, our lower FID shows CoMo-MUNIT better learned these complex visual transitions.

An alternative accuracy evaluation is proposed with a proxy task, which is an InceptionV3 network trained to regress sun elevation from real images and ϕ ground truths. For each method, we then generate 100 images at 100 ϕ locations, and measure the error between the input ϕ and the inference with the InceptionV3. Tab. 10b shows we outperform other methods with a $3.96°$ margin due to our better mapping.

4.5. Ablation studies

Architectural changes. We ablate the use of $\mathcal{L}_M$ and $\mathcal{L}_\phi$ by removing either. To evaluate the diversity of Day ↦ Timelapse translations, we sample 10 couples of random $\{\phi_1, \phi_2\}$ for 100 images and evaluate the LPIPS distance among translations pairs. We obtain LPIPS 0.020 *w/o* $\mathcal{L}_M$, 0.044 *w/o* $\mathcal{L}_\phi$, while using both proves best with **0.236**.

Disentangled reconstruction. While we disentangle real domain $Y(\phi)$ and model domain $Y_M(\phi)$ (cf. Fig. 2), steerable GANs [23] instead leverage guidance directly on $Y(\phi)$.

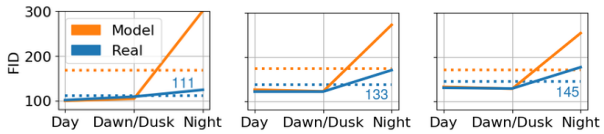


(a) Rolling FID

Method	Mean err. ↓	Std ↓
Model	21.12	10.15
DLOW [14]	17.39	9.02
StarGANV2 [9]	15.91	10.00
DNI-CycleGAN [64]	13.84	7.91
DNI-MUNIT [64]	13.80	8.30
CoMo-MUNIT	**9.84**	**7.20**
Real data	3.61	4.52

(b) ϕ regression

Figure 10: Evaluation of Day $\mapsto$ Timelapse. In a rolling FID (cf. text) shows our method is more effective in the complex dawn/dusk ("D/D") and night points, translating as lower mean FID (in legend). In b, we rank best on both mean and std error between the input ϕ and the regressed ϕ with an InceptionV3 network (trained on real data).



(a) Ours (b) [23] ($\lambda = 1$) (c) [23] ($\lambda = 5$)

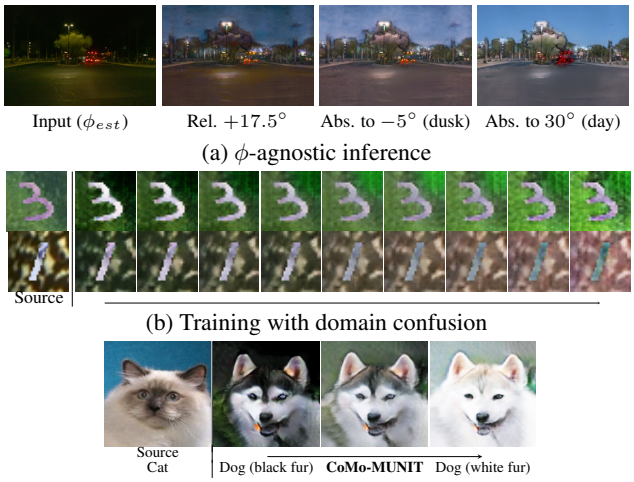
Figure 11: FIDs (cf. text) for *ours* (a) and steerable GANs [23] (b-c). *Ours* has lowest FIDs as it learns to depart from the model. Instead when increasing λ, [23] learns to mimic model but FID diverges from real images features.

To study either benefit, we replace $\mathcal{L}^{\phi}$ and $\mathcal{L}_M$ with $\mathcal{L}_{edit} = \lambda||y^{\phi} - \tilde{y}^{\phi}_M||_1$ as in [23]. Fig. 11 shows discrete FIDs, for *ours* and [23] with $\lambda = 1, 5$, evaluated against real data (blue) or model translations (orange). The plots hold complex but interesting insights. Specifically, low FIDs at Dawn/Dusk infer the model is reliable there, while divergent FIDs at night mean the opposite. With $\lambda = 1$ the i2i lacks guidance and performs poorly, but higher λ increases model mimicking and lower *real* FID. Instead, *ours* is *guided by the model* but learns to depart from it with the discovery of exclusive target features.

Model choice. We study the benefit of FIN encoding by swapping linear and cyclic. Comparing with Tab. 1, training iPhone $\mapsto$ DSLR with *cyclic FIN* is worse (IS/CIS/LPIPS 1.41/1.20/0.678) and at the cost of more complex encoding. Training Day $\mapsto$ Timelapse with *linear FIN* performs on par or better (IS/CIS/LPIPS 1.65/1.64/0.579) but *loses dusk/dawn distinction* capability.

5. Discussion

ϕ-agnostic inference. In all experiments, translation assumes source at ϕ_0, though agnostic inference is of interest. To test this, we trained our method with cycle consistency and shared parameters for $X \mapsto Y$ and $Y \mapsto X$ encoder/decoders (refer to Sec. 3.4). At inference, we used ϕ-Net$_A$ to estimate ϕ_{est} on input which enabled absolute translation regardless of input (e.g. anytime$\mapsto$day) but also relative translation (e.g. $+5°$). Sample results in Fig. 12a show exciting results with challenging night input.



(a) ϕ-agnostic inference

(b) Training with domain confusion

(c) Cat $\mapsto$ Dog with fur color guidance

Figure 12: a: Training with shared encoder/decoder and using ϕ-Net$_A$ at inference enables relative and absolute ϕ translations. The input is estimated at $\phi_{est} = -33.45°$ (gt $-32.73°$) and shifted with various strategies. b: CoMo-CycleGAN on MNIST-M [11] trained with *domain confusion* (w/o fixed ϕ), guiding on brightness (1st row) or redness (2nd). It shows source (leftmost) and translations along ϕ dimension. Despite domain confusion, it reorganized the manifold and produced valid translations. In c, we guide the complex Cat $\mapsto$ Dog only with fur color.

Source/Target domains confusion. A limitation of most GANs is the need of source/target splits while *truly unsupervised* GAN could discover a continuous manifold from mixed source/target data (i.e. $X \cup Y$ or domains confusion). Interestingly, model-guided GANs allow this *if* the model does not enforce ϕ input. While there are no physical model for bilateral night $\leftrightarrow$ day or foggy $\leftrightarrow$ clear, we prove the feasibility on MNIST-M [11] toy tasks, learning *brightness* or *redness* manifold. Fig. 12b shows we correctly achieve translation, paving ways for truly unsupervised GAN.

Models and data limitations. Model-guided GAN are unsuitable for some complex scenarios (e.g. face-to-face) due to the lack of models, but can guide features as skin tone, etc. as in our experiment Fig. 12c on Cat $\mapsto$ Dog using fur color guidance. Like [23], we too experienced that data scarcity affects greatly the manifold discovery and training timelapse without dusk and dawn proves to fail drastically.

References

[1] Pysolar. https://github.com/pingswept/pysolar. 4

[2] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *ICML*, 2018. 1

[3] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *CVPR*, 2020. 2

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

[5] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *CVPR*, 2020. 1

[6] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019. 1, 2

[7] Anoop Cherian and Alan Sullivan. Sem-gan: Semantically-consistent image-to-image translation. In *WACV*, 2019. 1

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 1

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1, 2, 6, 7, 8

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 4, 5, 6, 7

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 8

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. 2

[13] Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, and Luc Van Gool. Analogical image translation for fog generation. In *AAAI*, 2021. 1, 2

[14] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, 2019. 1, 2, 3, 6, 7, 8

[15] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *ICCV*, 2019. 2, 4, 5, 6, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[18] Judy Hoffman, Eric Tzeng, Taesung Park, and Phillip Isola Jun-Yan Zhu, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *ICML*, 2018. 1

[19] Lukas Hosek and Alexander Wilkie. An analytic model for full spectral sky-dome radiance. *TOG*, 2012. 4

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 2, 4, 5, 6

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[23] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *ICLR*, 2020. 2, 7, 8

[24] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 2

[25] A. Karpatne, William Watkins, J. Read, and V. Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv*, 2017. 2

[26] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *IJCV*, 2020. 2

[27] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *BMVC*, 2018. 1

[28] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *CVPR*, 2019. 2

[29] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 1

[30] Che-Tsung Lin, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Multimodal structure-consistent image-to-image translation. In *AAAI*, 2020. 1

[31] Jianxin Lin, Zhibo Chen, Yingce Xia, Sen Liu, Tao Qin, and Jiebo Luo. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. *T-PAMI*, 2019. 2

[32] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 1

[33] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 2

[34] Boyu Lu, Jun-Cheng Chen, and Rama Chellappa. Unsupervised domain-specific deblurring via disentangled representations. In *CVPR*, 2019. 2

[35] Björn Lütjens, Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodriguez, Océane

Boulais, Aaron Piña, Dava Newman, Alexander Lavin, Yarin Gal, et al. Physics-informed gans for coastal flood visualization. *arXiv*, 2020. 1

[36] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *ICLR*, 2019. 1

[37] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Siwei Ma, and Ming-Hsuan Yang. Continuous and diverse image-to-image translation via signed attribute vectors. *arXiv*, 2020. 2

[38] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 3

[39] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *ICLR*, 2019. 1

[40] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *CVPR*, 2018. 1

[41] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. In *BMVC*, 2020. 1

[42] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019. 2

[43] Jinshan Pan, Jiangxin Dong, Yang Liu, Jiawei Zhang, Jimmy Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang. Physics-based generative adversarial models for image restoration and beyond. *T-PAMI*, 2018. 2

[44] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 2

[45] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1

[46] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. Model-based occlusion disentanglement for image-to-image translation. In *ECCV*, 2020. 1, 2

[47] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *WACV*, 2020. 1, 6

[48] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *IJCV*, 2020. 1, 2

[49] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 2019. 2

[50] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *ICCV Workshops*, 2019. 2

[51] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 4, 5, 7

[52] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In *ECCV*, 2020. 2

[53] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018. 6

[54] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 5

[55] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *CVPR*, 2019. 1

[56] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019. 2

[57] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 4

[58] Hao Tang, Dan Xu, Yan Yan, Jason J Corso, Philip HS Torr, and Nicu Sebe. Multi-channel attention selection gans for guided image-to-image translation. In *CVPR*, 2019. 1

[59] William B Thompson, Peter Shirley, and James A Ferwerda. A spatial post-processing algorithm for images of night scenes. *Journal of Graphics Tools*, 2002. 4

[60] Marco Toldo, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for mobile semantic segmentation based on cycle consistency and feature alignment. *Image and Vision Computing*, 2020. 1

[61] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul de Charette, and Jean-François Lalonde. Rain rendering for evaluating and improving robustness to bad weather. *IJCV*, 2020. 1, 2

[62] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, 2017. 3

[63] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *CVPR*, 2017. 2

[64] Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. Deep network interpolation for continuous imagery effect transition. In *CVPR*, 2019. 1, 2, 6, 7, 8

[65] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *ICCV*, 2019. 1, 2

[66] Weihao Xia, Yujiu Yang, and Jing-Hao Xue. Unsupervised multi-domain multimodal image-to-image translation with explicit domain-constrained disentanglement. *Neural Networks*, 2020. 2

[67] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *ICLR Workshops*, 2018. 2

[68] You Xie, Erik Franz, Mengyu Chu, and Nils Thuerey. tempogan: A temporally coherent, volumetric gan for super-resolution fluid flow. *SIGGRAPH*, 2018. 2

[69] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *AAAI*, 2018. 2

[70] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 1

[71] Jianfu Zhang, Yuanyuan Huang, Yaoyi Li, Weijie Zhao, and Liqing Zhang. Multi-attribute transfer via disentangled representation. In *AAAI*, 2019. 2

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[73] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 6

[74] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. Forkgan: Seeing into the rainy night. In *ECCV*, 2020. 1

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017. 1, 4, 6, 7

[76] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 1

[77] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020. 1

[78] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *CVPR*, 2020. 1