Model-based occlusion disentanglement for image-to-image translation

Fabio Pizzati[1,2], Pietro Cerri[2], and Raoul de Charette[1⋆]

[1] Inria, Paris, France
{fabio.pizzati, raoul.de-charette}@inria.fr
[2] VisLab, Parma, Italy
pcerri@ambarella.com

Abstract. Image-to-image translation is affected by entanglement phenomena, which may occur in case of target data encompassing occlusions such as raindrops, dirt, etc. Our unsupervised model-based learning disentangles scene and occlusions, while benefiting from an adversarial pipeline to regress physical parameters of the occlusion model. The experiments demonstrate our method is able to handle varying types of occlusions and generate highly realistic translations, qualitatively and quantitatively outperforming the state-of-the-art on multiple datasets.

Keywords: GAN, image-to-image translation, occlusions, raindrop, soil

Fig. 1: Our method learns to disentangle scene from occlusions using unsupervised adversarial disentanglement with guided injection of a differentiable occlusion model. Here, we separate unfocused drops from rainy scene and show that, opposed to existing baselines, we learn a fully disentangled translation without drops and can re-inject occlusions with unseen parameters (e.g. in-focus drops).

1 Introduction

Image-to-image (i2i) translation GANs are able to learn the source $\mapsto$ target style mapping of paintings, photographs, etc. [54,20,16]. In particular, synthetic to real [4] or weather translation [29,38,27] attracted many works as they are alternatives to the menial labeling task, and allow domain adaptation or finetuning to boost performance on unlabeled domains. However, GANs notoriously fail

⋆ Corresponding author.

to learn the underlying physics [45]. This is evident when *target* data encompass occlusions (such as raindrop, dirt, etc.) as the network will learn an entangled representation of the scene with occlusions. For example, with clear $\mapsto$ rain the GAN translation will tend to have too many drops occlusions, often where the translation is complex as it is an easy way to fool the discriminator.

We propose an unsupervised model-based adversarial disentanglement to separate target and occlusions. Among other benefits, it enables accurate translation to the target domain and permits proper re-injection of occlusions. More importantly, occlusions with different physical parameters can be re-injected, which is crucial since the appearance of occlusions varies greatly with the camera setup. For example, drops occlusions appear different when imaged in-focus or out-of-focus. There are obvious benefits for occlusion-invariant outdoor vision like mobile robotics or autonomous driving. A comparison showcasing standard i2i (that partially entangles unrealistic drops) and our framework capabilities is available in Fig. 1. Our method builds on top of existing GAN architectures enabling unsupervised adversarial disentanglement with the only prior of the occlusion model. Parameters of the occlusion model are regressed on the target data and used when training to re-inject occlusions further driven by our disentanglement guide. We demonstrate our method is the only one able to learn an accurate translation in the context of occlusions, outperforming the literature on all tested metrics, and leading to better transfer learning on semantic segmentation. Our method is able to cope with various occlusion models such as drops, dirt, watermark, or else, among which raindrop is thoroughly studied. Our contributions may be summarized as follows:

- we propose the first unsupervised model-based disentanglement framework,
- our adversarial parameter estimation strategy allows estimating and replicating target occlusions with great precision and physical realism,
- our disentanglement guidance helps the learning process without losing generative capabilities in the translation task,
- we conducted exhaustive experiments on raindrops occlusions proving we outperform the literature, boost transfer learning, and provide a focus agnostic framework of high interest for autonomous driving applications.

2 Related work

Image-to-image translation. Seminal works on image-to-image translation (i2i) was conducted by Isola et al. [16] and Zhu et al. [54] for paired and unpaired data respectively, where the later introduced the cycle consistency loss extended in [55,48]. Liu et al. [20] further proposed using Variational Auto Encoders to learn a shared latent space. A common practice to increase accuracy is to learn scene-aware translation exploiting additional supervision from semantic [19,30,41,6], instance [25] or objects [38]. Furthermore, a recent trend is to use attention-guided translation [24,22,40,17] to preserve important features.
Regarding disentangled representations, MUNIT [14] and DRIT [18] decouple image content and style to enable multi-modal i2i, similar in spirit to our goal,

while FUNIT [21] uses disentangled representations for few-shot learning. FineGAN [39] disentangles background and foreground using bounding boxes supervision. Multi-domain i2i also opens new directions to control elements at the image level [7,32,3,47,15], since it may be used to represent elements learned from various datasets. Attribute-based image generation [43,44,51] follows a similar scheme, explicitly controlling features. Nonetheless, these methods require attributes annotations or multiple datasets – hardly compatible with occlusions. Finally, Yang et al. [46] exploit a disentangled physical model for dehazing.

Lens occlusion generation (drops, dirt, soiling, etc.). Two strategies co-exist in the literature: physics-based rendering or generative networks. Early works on geometrical modeling showcased accurate rendering of raindrops via ray-tracing and 3D surface modeling [34,35,13], sometimes accounting for complex liquid dynamics [50] or focus blur [35,13]. A general photometric model was also proposed in [10] for thin occluders, while recent works use displacement maps to approximate the raindrops refraction behavior [28,2]. Generative networks were also recently leveraged to learn general dirt generation [42] but using semantic soiling annotations. To the best of our knowledge, there are no approaches that simultaneously handle occlusions and scene-based modifications with i2i. Note that we intentionally do not review the exhaustive list of works on de-raining or equivalent as it is quite different from disentanglement in the i2i context.

3 Model-based disentanglement

We aim to learn the disentangled representation of a target domain and occlusions. For example, when translating clear to rain images having raindrops on the lens, standard image-to-image (i2i) fails to learn an accurate mapping as the target entangles the scene and the drops on the lens. We depart from the literature by learning a disentangled representation of the target domain from the injection of an occlusion model, in which physical parameters are estimated from the target dataset. Not only does it allows us to learn the disentangled representation of the scene (e.g. target image without any occlusions) but also to re-inject the occlusion model either with the estimated parameters or with different parameters (e.g. target image with drops in focus).

Our method handles any sort of occlusions such as raindrops, soil, dirt, watermark, etc. but for clarity we focus on raindrops as it of high interest and exhibits complex visual appearance. Fig. 2 shows an overview of our training pipeline, which is fully unsupervised and only exploits the prior of occlusion type. To learn adversarial disentanglement by injecting occlusions (Sec. 3.1), we first pretrain a baseline to regress the model parameters (Sec. 3.2) and estimate domain shifts to further guide the disentanglement learning process (Sec. 3.3).

3.1 Adversarial disentanglement

Let X and Y be the domains of a *source* and a *target* dataset, respectively. In an i2i setup, the task is to learn the $X \mapsto Y$ mapping translating source to

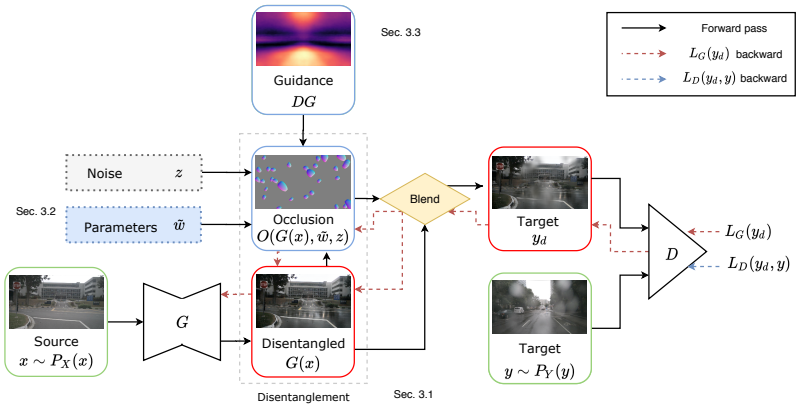


Fig. 2: To disentangle the i2i translation process in an unsupervised manner, we inject occlusions $O(.)$ with estimated parameters $\tilde{w}$ before forwarding the generated image $G(x)$ through the discriminator D. The Disentanglement Guidance (DG) avoids losing translation capabilities in low domain shift areas. *Fake* and *real* data are drawn red and green, respectively.

target. Now, if the target dataset has occlusions of any sort, Y encompasses two domains: the scene domain $_S$, and the occlusion domain $_O$. Formally, as in [27] we introduce a disentangled representation of domains such that $Y = \{Y_S, Y_O\}$ and $X = \{X_S\}$. In adversarial training strategies, the generator is led to approximate P_X and P_Y, the probability distributions associated with the domains stochastic process, defined as

$$\begin{aligned} \forall x \in X, x \sim P_X(x), \\ \forall y \in Y, y \sim P_Y(y). \end{aligned} \tag{1}$$

Having occlusions, the target domain Y is interpreted as the composition of two subdomains, and we seek to estimate $P_{Y_S,Y_O}(y_S, y_O)$ corresponding to the scene and occlusion domain. To address this, let's make the *naive* assumption that marginals $P_{Y_S}(y_S)$ and $P_{Y_O}(y_O)$ are independent from each other. Thus, exploiting the definition of joint probability distribution, $P_Y(y)$ becomes

$$P_Y(y) = P_{Y_S,Y_O}(y_S, y_O) = P_{Y_S}(y_S)P_{Y_O}(y_O), \tag{2}$$

and it appears that knowing one of the marginals would enable learning disentangled i2i translations between subdomains. In particular, if $P_{Y_O}(y_O)$ is known it is intuitively possible to learn $X_S \mapsto Y_S$, satisfying our initial requirement.

In reality, transparent occlusions - such as raindrops - are not fully disentangled from the scene, as their appearance is varying with scene content (see ablation in Sec. 4.4). However, the physical properties of occlusions may be seen as quite independent. As an example, while the appearance of drops on the lens varies greatly with scene background, their physics (e.g. size, shape, etc.) is little-

or un- related to the scene. Fortunately, there is extensive literature providing appearance models for different types of occlusions (drop, dirt, etc.) given their physical parameters, which we use to estimate $P_{Y_O}(y_O)$. We thus formalize occlusion models as $O(s, w, z)$ parametrized by the scene s, the set of disentangled physical properties w, and a random noise vector z. The latter is used to map characteristics that can not be regressed as they are stochastic in nature. This is the case for raindrops positions for example. Assuming we know the type of occlusion such as drop, dirt, etc., we rely on existing models (described in Sec. 4) to render the visual appearance of occlusions.

Ultimately, as depicted in Fig. 2, we add occlusions rendered with a known model on generated images before forwarding them to the discriminator. Assuming sufficiently accurate occlusion models, the generator G is pushed to estimate the disentangled P_{Y_S}, thus learning to translate only scene-related features. As a comparison to a standard LSGAN [23] training which enforces a zero-sum game by minimizing

$$\begin{aligned} y_{\mathrm{d}} &= G(x), \\ L_{\mathrm{gen}} &= L_G(y_d) = \mathbb{E}_{x\ \sim P_X(x)}[(D(y_d) - 1)^2], \\ L_{\mathrm{disc}} &= L_D(y_d, y) = \mathbb{E}_{x\ \sim P_X(x)}[(D(y_d))^2] + \mathbb{E}_{y\ \sim P_Y(y)}[(D(y) - 1)^2], \end{aligned} \tag{3}$$

with L_{gen} and L_{disc} being respectively the tasks of the generator G and discriminator D, we instead learn the desired disentangled mapping by injecting occlusions $O(.)$ on the translated image. Hence, we newly define y_{d} as the disentangled composition of translated scene and injected occlusions, that is

$$y_{\mathrm{d}} = \alpha G(x) + (1 - \alpha)O(G(x), \tilde{w}, z), \tag{4}$$

where α is a pixel-wise measure of the occlusion transparency. Opaque occlusions will locally set $\alpha = 1$ while a pixel with transparent occlusions has $\alpha < 1$. Because physical parameters greatly influence the appearance of occlusions (e.g. drop in focus or out of focus), we render occlusions in Eq. 4 using $\tilde{w}$, the optimal set of physical parameters to model occlusions in Y.

3.2 Adversarial parameters estimation

We estimate the set of physical parameters $\tilde{w}$ from Y in an unsupervised manner, benefiting from the entanglement of scene and occlusion in the target domain. We build here upon Eq. 2. Assuming a naive i2i baseline being trained on source and target data, *without* disentanglement, the discriminator learned to distinguished examples from source and target by discriminating $P_X = P_{X_S}$ from $P_Y = P_{Y_S}(y_S)P_{Y_O}(y_O)$. Let's consider the trivial case where a generator G' performs identity (i.e. $G'(x) = x$) then we get $P_{Y_S}(y_S) = P_{X_S}(x_S)$ and it is now possible to estimate the optimal $\tilde{w}$ as it corresponds to the distance minimization of P_Y and P_X. Intuitively, as the domain gap results of both occlusion and

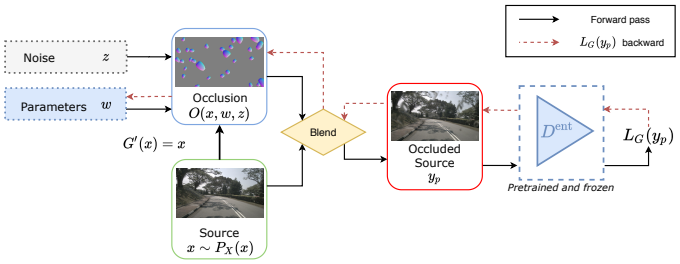


Fig. 3: We estimate the optimal parameters to use for the disentanglement adding occlusions on *source* images and optimizing the parameters of the physical model in order to fool the discriminator. Since we are not using a generator network, the gradient, represented as red arrows, flows only in the occlusion model direction.

scene domain (which is fixed), reducing the source and target domain gap implies reducing the occlusion domain gap, in extenso regressing w.

Fig. 3 illustrates the estimation process. In practice, we pretrain a simple i2i baseline (e.g. MUNIT[14]) to learn $X \mapsto Y$ in a naive - entangled - manner, by training a generator and discriminator. We then freeze the naive discriminator denoted D^{ent} and solve the following optimization objective

$$\begin{gathered} y_p = \alpha G'(x) + (1-\alpha) O(G'(x), w, z), \\ \min_w L_G(y_p)\,, \end{gathered} \tag{5}$$

by backpropagating the gradient flow through the derivable occlusion model. For most occlusions where transparency depends on the model, we in fact consider the blending mask $\alpha = \alpha(w, z)$. Note that it is required to freeze the discriminator otherwise we would lose any feedback capabilities on the images of the target domain. For simplicity in Fig. 3, we omit the generator G' during parameter estimation since $G'(x) = x$. Training until convergence, we extract the optimal parameter set $\tilde{w}$. In Sec. 4.2 we evaluate our parameters estimation on synthetic and real data.

Alternately, $\tilde{w}$ could also be tuned manually but at the cost of menial labor and obvious approximation. Still, one may note than an inaccurate estimation of $\tilde{w}$ would lead to a poor disentanglement of P_{Y_S} and P_{Y_O}.

3.3 Disentanglement guidance

We highlight now an easy pitfall in the disentangled GAN training, since an unwanted optimum is reached if the generator simply adds occlusions. Indeed because occlusions are visually simple and constitute a strong discriminative signal for the discriminator, it may be easier for the generator to entangle occlusions rather than to learn the underlying scene mapping. Specifically, occlusions will be entangled where source and target differ the least, since it is an easy way to minimize L_{gen} as even with a perfect i2i *there* the discriminator will provide

relatively uncertain feedback. For example, we noticed that drops were entangled over trees or buildings as both exhibit little visual differences in the clear and rainy domains.

To avoid such undesirable behavior, we spatially guide the disentanglement to prevent the i2i task from entangling occlusions in the scene representation. The so-called *disentanglement guide* is computed through the estimation of the domains gap database-wide. Specifically, we use GradCAM [37] which relies on gradient flow through the discriminator to identify which regions contribute to the *fake* classification, and thus exhibit a large domain gap. Similar to the parameter estimation (Sec. 3.2), we pretrain a simple i2i baseline and exploit the discriminator. To preserve resolution, we upscale and average the response of GradCAM for each discriminator layer and further average responses over the dataset[3]. Formally, using LSGAN we extract *Disentanglement Guidance (DG)*

$$DG = \mathbb{E}_{x\ \sim P_X(x)}[\mathbb{E}_{l\in L}[\mathrm{GradCAM}_l(D(x))]\,, \tag{6}$$

with L being the discriminator layers. During training of our method, the guide serves to inject occlusions only where domain gaps are low, that is where $DG < \beta$, with $\beta \in [0, 1]$ a hyperparameter. While this may seem counter-intuitive, explicitly injecting drops in low domain shift areas mimics the GAN intended behavior lowering the domain shift with drops. This logically prevents entanglement phenomena since they are simulated by the injection of occlusions. We refer to Fig. 8b in the ablation study for a visual understanding of this phenomenon.

4 Experiments

We validate the performances of our method on various real occlusions, leveraging recent real datasets such as nuScenes [5], RobotCar [28], Cityscapes [9] or WoodScape [49], and synthetic data such as Synthia [33]. Our most comprehensive results focus on the harder raindrops occlusion (Sec. 4.2), but we also extend to soil/dirt and other general occlusions such as watermark, fence, etc. (Sec. 4.3). For each type of occlusion we detail the model used and report qualitative and quantitative results against recent works: DRIT [18], U-GAT-IT [17], AttentionGAN [40], CycleGAN [54], and MUNIT [14]. Because the literature does not account for disentanglement, we report both the disentangled underlying domain (*Ours disentangled*) and the disentangled domain *with* injection of *target* occlusions (*Ours target*). Note that while still images are already significantly better with our method, the full extent is better sensed on the supplementary video as disentangling domains implicitly enforces temporal consistency.

4.1 Training setup

Our method is trained in a three stages unsupervised fashion, with the only prior that the occlusion model is known (e.g. drop, dirt, watermark, etc.). First, we

[3] Note that averaging through the dataset implies similar image aspects and viewpoints. Image-wise guidance could be envisaged at the cost of less reliable guidance.

train an i2i baseline to learn the entangled source $\mapsto$ target and extract D^{ent}. Second, the occlusion model parameters are regressed as in Sec. 3.2 and DG is estimated with the same pre-trained discriminator following Sec. 3.3. While being not mandatory for disentanglement, DG often improves visual results. Third, the disentangled pipeline described in Sec. 3.1 is trained from scratch injecting the occlusion model only where the Disentanglement Guidance allows it. We refer to the supplementary for more details.
We use MUNIT [14] for its multi-modal capacity and train with LSGAN [23]. Occlusion models are implemented in a differentiable manner with Kornia [31].

4.2 Raindrops

We now evaluate our method on the complex task of raindrops disentanglement when learning the i2i clear $\mapsto$ rain task. Because of their refractive and semi-transparent appearance, raindrops occlusions are fairly complex.

Occlusion model. To model raindrops, we use the recent model of Alletto *et al.* [2] which provides a good realism/simplicity trade-off. Following [2], we approximate drop shapes with simple trigonometric functions and add random noise to increase variability as in [1]. The photometry of drops is approached with a displacement map (U, V) encoding the 2D coordinate mapping in the target image, such that drop at (u, v) with thickness ρ has its pixel (u_i, v_i) mapped to

$$\big(u + \mathrm{U}(u_i, v_i) \cdot \rho, v + \mathrm{V}(u_i, v_i) \cdot \rho\big)\,. \tag{7}$$

Intuitively, this approximates light refractive properties of raindrops. Technically, $(\mathrm{U}, \mathrm{V}, \rho)$ is conveniently encoded as a 3-channels image. We refer to [2] for details. We also account for the imaging focus, as it has been highlighted that drops with different focus have dramatically different appearance [12,8,2]. We approximate focus blur with a Gaussian point spread function [26] which variance σ is learned, thus $w = \{\sigma\}$. I.e., our method handles drops occlusions with any type of focus. During training, drops are uniformly distributed in the image space, with size being a hyperparameter which we study later, and defocus σ is regressed with our parameters estimation (Sec. 3.2). During inference, drops are generated at random position with p_{r} probability, which somehow controls the rain intensity. Fig. 4 illustrates our drop occlusion model with variable shapes and focus blur.

Datasets. We evaluate using 2 recent datasets providing clear/rain images.
nuScenes [5] is an urban driving dataset recorded in the US and Singapore with coarse frame-wise weather annotation. Using the latter, we split the validation into clear/rain and obtain 114251/29463 for training and 25798/5637 for testing.
RobotCar [28] provides pairs of clear/rain images acquired with binocular specialized hardware where one camera is continuously sprayed with water. The clear images are warped to the rainy image space using calibration data, and we use a clear/rain split of 3816/3816 for training and 1000/1000 for validation.

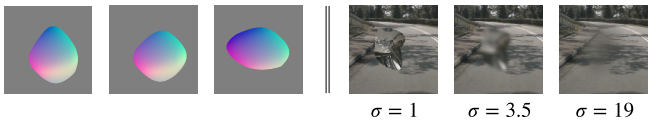


Fig. 4: Raindrop occlusion model. Left are schematic views of our model where the shape is modeled as trigonometric functions and photometry as displacement maps (cf. Eq. 7), encoded here as RGB. Right demonstrates our ability to handle the high variability of drops appearance with a different focus (σ).

Qualitative evaluation. Outputs of the clear $\mapsto$ rain i2i task are shown in Fig. 5 against the above cited [18,17,40,54,14]. At first sight, it is evident that drops are entangled in other methods, which is expected as they are *not* taught to disentangle drops. To allow fair comparison we thus provide our disentangled estimation (*Ours disentangled*) but also add drops occlusions to it, modeled with the physical parameters $\tilde{w}$ estimated from target domain (*Ours target*).

Looking at *Ours disentangled*, the i2i successfully learned the appearance of a rainy scene (e.g. reflections or sky) with sharp pleasant translation to rain, without any drops. Other methods noticeably entangle drops, often at fixed positions to avoid learning i2i translation. This is easily noticed in the 4th column where all methods generated drops on the leftmost tree. Conversely, we benefit from our disentangled representation to render scenes with drops occlusions (row *Ours target*) fairly matching the appearance of the target domain (1st row), subsequently demonstrating the efficiency of our adversarial parameter estimation.

What is more, we inject drops with different sets of parameters $\{w, z\}$ arbitrary mimicking dashcam sequences (Fig. 5, last 2 rows). The quality of the dashcam translations, despite the absence of similar data during training, proves the benefit of disentanglement and the adequacy of the occlusion model. Note that with any set of parameters, our occlusion (last 3 rows) respect the refractive properties of raindrops showing the scene up-side-down in each drop, while other baselines simply model white and blurry occlusions.

Quantitative evaluation.

GAN metrics. Tab. 1a reports metrics on the nuScenes clear $\mapsto$ rain task. Each metric encompasses different meanings: Inception Score (IS) [36] evaluates quality/diversity against target, LPIPS distance [52] evaluates translation diversity thus avoiding mode-collapse, and Conditional Inception Score [14] single-image translations diversity for multi-modal baselines. Note that we evaluate against our disentangled + target drops occlusion *Ours target*, since baselines are neither supposed to disentangle the occlusion layer nor to generate different kinds of drop. On all metrics, our method outperforms the state of the art by a comfortable margin. This is easily ascribable to our output being both more realistic, since we are evaluating with drops with the physical parameters extracted from target dataset, and more variable, since we do not suffer from entanglement

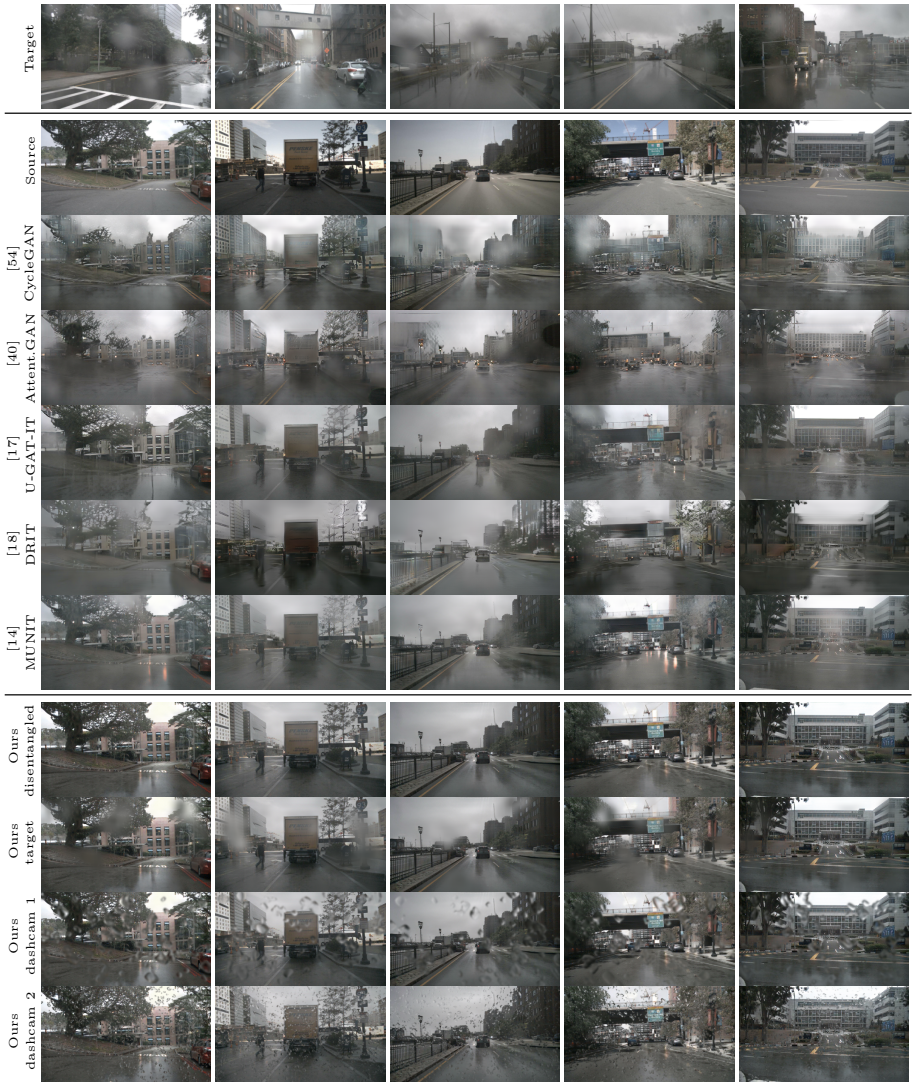


Fig. 5: Qualitative comparison against recent baselines on the clear ↦ rain task with drops occlusions. Target samples are displayed in the first row for reference. Other rows show the source image (2nd row) and all its subsequent translations below. Our method efficiently disentangled drops occlusion from the scene (row *Ours disentangled*) and subsequently allows the generation of realistic drops matching target style (row *Ours target*) or any arbitrary style (last 2 rows).

Network	IS↑	LPIPS↑	CIS↑
CycleGAN [54]	1.151	0.473	-
AttentionGAN [40]	1.406	0.464	-
U-GAT-IT [17]	1.038	0.489	-
DRIT [18]	1.189	0.492	1.120
MUNIT [14]	1.211	0.495	1.030
Ours target	**1.532**	**0.515**	**1.148**

(a) GAN metrics

Method	AP↑
Original (from [11])	18.7
Finetuned w/ Halder *et al.* [11]	25.6
Finetuned w/ Ours target	**27.7**

(b) Semantic segmentation

Table 1: Quantitative evaluation of clear ↦ rain effectiveness on nuScenes [5] (for all higher is better). (a) shows GAN metrics of ours i2i translation with target drop inclusion (i.e. *Ours Target*) against i2i baselines. Our method outperforms literature on all metrics which is imputed to the variability and realism that come with the disentanglement. Note that CIS is multi-modal. (b) Evaluation of the Average Precision (AP) of semantic segmentation when finetuning PSPNet [53] and evaluating on a subset of nuScenes with semantic labels from [11].

phenomena that greatly limit the drops visual stochasticity. This is also evident when comparing against [14] which we use as the backbone in our framework. Technically, IS is computed over the whole validation set, CIS on 100 translations of 100 random images (as in [14]), and LPIPS on 1900 random pairs of 100 translations. The InceptionV3 for IS/CIS was finetuned on source/target as [14].

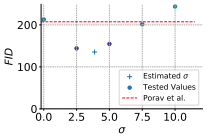
Semantic segmentation. Because GAN metrics are reportedly noisy [52] we aim at providing a different perspective for quantitative evaluation, and thus measure the usefulness of our translated images for semantic segmentation. To that aim, following the practice of Halder et al. [11] we use *Ours target* (i.e. trained on rainy nuScenes) to infer a rainy version of the popular Cityscapes [9] dataset and use it to finetune PSPNet [53]. The evaluation on the small subset of 25 semantic labeled images of rainy nuScenes provided by [11] is reported in Tab. 1b. It showcases finetuning with our rainy images is better than with [11], which uses physics-based rendering to generate rain. Note that both finetune *Original* weights, and that the low numbers results of the fairly large Cityscapes-nuScenes gap (recall that nuScenes has no semantic labels to train on).

Parameter estimation. We verify the validity of our parameter estimation strategy (Sec. 3.2) using the RobotCar dataset, which provides real *clear/rain* pairs of images. As the viewpoints are warped together (cf. Datasets details above), there is no underlying domain shift in the clear/rain images so we set $G(x) = x$ and directly train on discriminator to regress physical parameters (we get $\sigma = 3.87$) and render rain with it on clear images. We can then measure the distance of the translated and real rainy images, with FID and LPIPS distances reported in Fig. 6b. Unlike before, LPIPS measures distance (not diversity) so lower is better. For both we significantly outperform [28], which is visually inter-

(a) Sample images

Method	FID↓	LPIPS↓
Porav et al. [28]	207.34	0.53
Ours target	**135.32**	**0.44**

(b) Benchmark on [28]



(c) FID

Fig. 6: Parameter estimation using real clear/rainy pairs of images from RobotCar [28]. Visually, *Ours target* is fairly closer to the *Target* sample regardless of drops position/size (a), while quantitatively lower FID and LPIPS distance is obtained (b). With FID measures at different defocus sigma in (c), we demonstrate our estimated parameters ($\sigma = 3.87$) successfully led to the best parameters.

pretable in Fig. 6a, where drops rendered with our parameter estimation looks more similar to *Target* than those of [28] (regardless of their size/position).

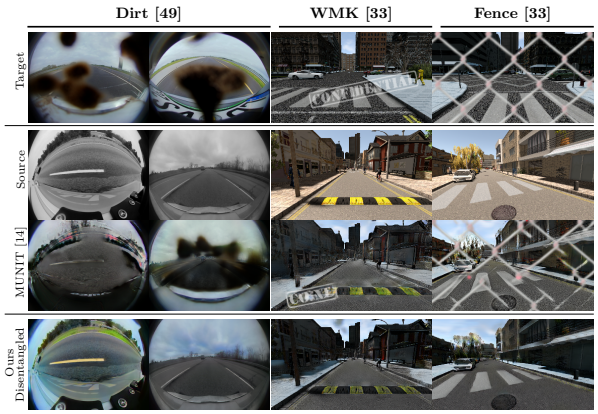
To further assess the accuracy of our estimation, we plot in Fig. 6c the FID for different defocus blurs ($\sigma \in \{0.0, 2.5, 5.0, 7.5, 10\}$). It shows our estimated defocus ($\sigma = 3.87$) leads to the minimum FID of all tested values, further demonstrating the accuracy of our adversarial estimation. We quantify the precision by training on clear images with synthetic injection of drops having $\sigma \in [5, 25]$, and we measured an average error of 1.02% (std. 1.95%).

4.3 Extension to other occlusion models

To showcase the generality of our method, we demonstrate its performance on two generally encountered types of occlusions: Dirt and General occlusion.

Dirt. We rely on the recent WoodScape dataset [49] and a simple occlusion model to learn the disentangled representation.
Datasets. WoodScape [49] provides a large amount of driving fish-eye images, and comes with metadata indicating the presence of dirt/soil[4]. Different from rain sequences having rainy scenes+drops occlusions, apart from soiling there isn't any domain shift in the clean/dirt images provided. Hence, to study disentanglement we introduce an additional shift by converting clean images to grayscale and refer to them as *clean_gray*. We train our method on non-paired clean_gray/dirt images with 5117/4873 for training and 500/500 for validation.
Occlusion model. To generate synthetic soiling, we use a modified version of our drop model with random trigonometric functions and larger varying sizes. Displacement maps are not used since we consider dirt to be opaque and randomly brownish, with apparent semi-transparency only as a result of the high defocus. As for drops, the defocus σ is regressed so again $w = \{\sigma\}$.
Performance. Fig. 7a (left) shows sample results where the task consists of disentangling the color characteristics from the dirt occlusion (since color is only in the

[4] Note that WoodScape provides soiling mask which we do *not* use.



(a) Qualitative results

Model	Network	IS↑	LPIPS↑	CIS↑
Dirt	MUNIT [14]	1.06	**0.66**	1.08
	Ours target	**1.26**	0.59	**1.15**
Fence	MUNIT [14]	1.26	**0.55**	1.11
	Ours target	**1.31**	0.54	**1.19**
WMK	MUNIT [14]	1.17	**0.57**	1.01
	Ours target	**1.19**	0.55	**1.02**

(b) GAN metrics

Model	Network	SSIM↑	PSNR↑
Dirt	MUNIT [14]	0.41	13.40
	Ours disent.	**0.76**	**20.23**

(c) Colorization metrics

Fig. 7: Various occlusions disentanglement. We seek to learn disentangled representation of clean_gray ↦ clean_color on WoodScape [49] (real) and clear ↦ snow on Synthia [33] (synthetic). For all, MUNIT [14] partly entangles occlusions in the translation, often occluding hard-to-translate areas, while our method learned correctly the color mapping and the snow mapping despite complex occlusions (7a). Quantitative evaluation with GAN metrics (7b) confirms the increase in image quality for all occlusions models and with colorization metrics for dirt (7c) exploiting our unpaired disentanglement framework.

dirt data). Comparing to MUNIT [14], *Ours disentangled* successfully learned color without dirt entanglement, while [14] failed to learn accurate colorization due to entanglement. Performances are validated quantitatively in Tab. 7b-7c.

General occlusions (synthetic). In Fig. 7a (right) we also demonstrate the ability to disentangle general occlusion, in the sense of an alpha-blended layer on an image (watermarks, logos, etc.). We used synthetic Synthia [33] clear/snow data, and augmented only snow either with a "confidential" watermark (WMK) or a fence image, both randomly shifted. Our i2i takes 3634/3739 clear/snow images for training, and 901/947 for validation. The occlusion model is the ground truth composite alpha-blended model, with random translation, and without any regressed parameters (i.e. $w = \emptyset$). From Fig. 7a, our method learned a disentangled representation, while MUNIT [14] partially entangled the occlusion model. In tab. 7b, CIS/IS confirm the higher quality visual results.

4.4 Ablation studies

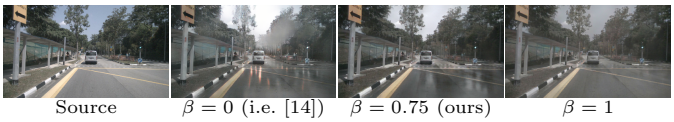
Model complexity. We study here how much model complexity impacts disentanglement, with the evaluation on the nuScenes clear ↦ rain task. We compare three decreasingly complex occlusion models: 1) *Ours*, the raindrop model de-

scribed in Sec. 4.2; 2) *Refract*, which is our model without any shape or thickness variability; 3) *Gaussian*, where drops are modeled as scene-independent Gaussian-shaped occlusion maps following [10]. From Fig. 8a, while *Ours* has best performance, even simpler models lead to better image translation which we relate to our disentanglement capability. To also assess that the occlusion model doesn't only play the role of an adversarial attack, we also compare the FID of real RobotCar raindrops (as in Sec. 4.2) when training with either of the models described in Sec. 4.2 and 4.3. The FID measured are **135.32** (drop) / 329.17 (watermark) / 334.76 (dirt) / 948.71 (fence). This advocates that *a priori* knowledge of the occlusion type is necessary to achieve good results.

Disentanglement Guidance (DG). We study the effects of guidance (eq. 6) on the nuScenes clear $\mapsto$ rain task, by varying the β threshold used to inject occlusion where $DG < \beta$. From Fig. 8b, with conservative guidance ($\beta = 0$, i.e. no occlusions injected) it behaves similar to MUNIT baseline entangling drops in the translation, while deactivating guidance ($\beta = 1$) correctly achieves a disentangled representation but at the cost of losing translation in high domain shifts areas (note the lack of road reflections). Appropriate guidance ($\beta = 0.75$) helps learning target characteristics while preserving from entanglement.

Model	IS↑	LPIPS↑	CIS↑
N/A ([14])	1.21	0.50	1.03
Gaussian	1.35	0.51	1.13
Refract	1.46	0.50	1.12
Ours	**1.53**	**0.52**	**1.15**

(a) Model complexity



(b) Disentanglement guidance

Fig. 8: Ablation of model complexity and disentanglement guidance for the clear $\mapsto$ rain task on nuScenes. In (a), our disentanglement performs better than baseline [14] with all occlusion models. In (b), studying the influence of β we note that without guidance ($\beta = 1$) the translation lacks important rainy features (reflections, glares, etc.) while with appropriate guidance ($\beta = 0.75$) it learns correct rainy characteristics without entanglement.

5 Conclusion

We propose the first unsupervised method for model-based disentanglement in i2i translation, relying on guided injection of occlusions with parameters regressed from target and assuming only prior knowledge of the occlusion model. Our method outperformed the literature visually and on all tested metrics, and the applicability was shown on various occlusions models (raindrop, dirt, watermark, etc.). Our strategy of adversarial parameter estimation copes with drops of any focus, which is of high interest for any outdoor system as demonstrated in the experiments.

References

1. Rain drops on screen. `https://www.shadertoy.com/view/ldSBWW`
2. Alletto, S., Carlin, C., Rigazio, L., Ishii, Y., Tsukizawa, S.: Adherent raindrop removal with self-supervised attention maps and spatio-temporal generative adversarial networks. In: ICCV Workshops (2019)
3. Anoosheh, A., Agustsson, E., Timofte, R., Van Gool, L.: Combogan: Unrestrained scalability for image domain translation. In: CVPR Workshops (2018)
4. Bi, S., Sunkavalli, K., Perazzi, F., Shechtman, E., Kim, V.G., Ramamoorthi, R.: Deep cg2real: Synthetic-to-real translation via image disentanglement. In: ICCV (2019)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
6. Cherian, A., Sullivan, A.: Sem-gan: Semantically-consistent image-to-image translation. In: WACV (2019)
7. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
8. Cord, A., Aubert, D.: Towards rain detection through use of in-vehicle multipurpose cameras. In: IV (2011)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
10. Gu, J., Ramamoorthi, R., Belhumeur, P., Nayar, S.: Removing image artifacts due to dirty camera lenses and thin occluders. In: SIGGRAPH Asia (2009)
11. Halder, S.S., Lalonde, J.F., de Charette, R.: Physics-based rendering for improving robustness to rain. In: ICCV (2019)
12. Halimeh, J.C., Roser, M.: Raindrop detection on car windshields using geometric-photometric environment construction and intensity-based correlation. In: IV (2009)
13. Hao, Z., You, S., Li, Y., Li, K., Lu, F.: Learning from synthetic photorealistic raindrop for single image raindrop removal. In: ICCV Workshops (2019)
14. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
15. Hui, L., Li, X., Chen, J., He, H., Yang, J.: Unsupervised multi-domain image translation with domain-specific encoders/decoders. In: ICPR (2018)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
17. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: ICLR (2020)
18. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. arXiv preprint arXiv:1905.01270 (2019)
19. Li, P., Liang, X., Jia, D., Xing, E.P.: Semantic-aware grad-gan for virtual-to-real urban scene adaption. BMVC (2018)
20. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NeurIPS (2017)

21. 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)
22. Ma, S., Fu, J., Wen Chen, C., Mei, T.: Da-gan: Instance-level image translation by deep attention generative adversarial networks. In: CVPR (2018)
23. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV (2017)
24. Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image-to-image translation. In: NeurIPS (2018)
25. Mo, S., Cho, M., Shin, J.: Instagan: Instance-aware image-to-image translation. ICLR (2019)
26. Pentland, A.P.: A new sense for depth of field. T-PAMI (1987)
27. Pizzati, F., de Charette, R., Zaccaria, M., Cerri, P.: Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In: WACV (2020)
28. Porav, H., Bruls, T., Newman, P.: I can see clearly now: Image restoration via de-raining. In: ICRA (2019)
29. Qu, Y., Chen, Y., Huang, J., Xie, Y.: Enhanced pix2pix dehazing network. In: CVPR (2019)
30. Ramirez, P.Z., Tonioni, A., Di Stefano, L.: Exploiting semantics in adversarial training for image-level domain adaptation. In: IPAS (2018)
31. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: WACV (2020)
32. Romero, A., Arbeláez, P., Van Gool, L., Timofte, R.: Smit: Stochastic multi-label image-to-image translation. In: ICCV Workshops (2019)
33. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
34. Roser, M., Geiger, A.: Video-based raindrop detection for improved image registration. In: ICCV Workshops (2009)
35. Roser, M., Kurz, J., Geiger, A.: Realistic modeling of water droplets for monocular adherent raindrop recognition using bezier curves. In: ACCV (2010)
36. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016)
37. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
38. Shen, Z., Huang, M., Shi, J., Xue, X., Huang, T.S.: Towards instance-level image-to-image translation. In: CVPR (2019)
39. Singh, K.K., Ojha, U., Lee, Y.J.: Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In: CVPR (2019)
40. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: International Joint Conference on Neural Networks (IJCNN) (2019)
41. Tang, H., Xu, D., Yan, Y., Corso, J.J., Torr, P.H., Sebe, N.: Multi-channel attention selection gans for guided image-to-image translation. In: CVPR (2019)
42. Uricar, M., Sistu, G., Rashed, H., Vobecky, A., Krizek, P., Burger, F., Yogamani, S.: Let's get dirty: Gan based data augmentation for soiling and adverse weather classification in autonomous driving. arXiv preprint arXiv:1912.02249 (2019)
43. Xiao, T., Hong, J., Ma, J.: Dna-gan: Learning disentangled representations from multi-attribute images. ICLR Workshops (2018)

44. Xiao, T., Hong, J., Ma, J.: Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: ECCV (2018)
45. Xie, Y., Franz, E., Chu, M., Thuerey, N.: tempogan: A temporally coherent, volumetric gan for super-resolution fluid flow. SIGGRAPH (2018)
46. Yang, X., Xu, Z., Luo, J.: Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In: AAAI (2018)
47. Yang, X., Xie, D., Wang, X.: Crossing-domain generative adversarial networks for unsupervised multi-domain image-to-image translation. In: MM (2018)
48. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
49. Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricár, M., Milz, S., Simon, M., Amende, K., et al.: Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In: ICCV (2019)
50. You, S., Tan, R.T., Kawakami, R., Mukaigawa, Y., Ikeuchi, K.: Adherent raindrop modeling, detectionand removal in video. T-PAMI (2015)
51. Zhang, J., Huang, Y., Li, Y., Zhao, W., Zhang, L.: Multi-attribute transfer via disentangled representation. In: AAAI (2019)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
53. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
54. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: CVPR (2017)
55. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NeurIPS (2017)