Look ATME: The Discriminator Mean Entropy Needs Attention

Edgardo Solano-Carrillo Angel Bueno Rodriguez Borja Carrillo-Perez
Yannik Steiniger Jannis Stoppe
German Aerospace Center (DLR), Institute for the Protection of Maritime Infrastructures
{Edgardo.SolanoCarrillo, Angel.Bueno, Borja.CarrilloPerez, Yannik.Steiniger, Jannis.Stoppe}@dlr.de

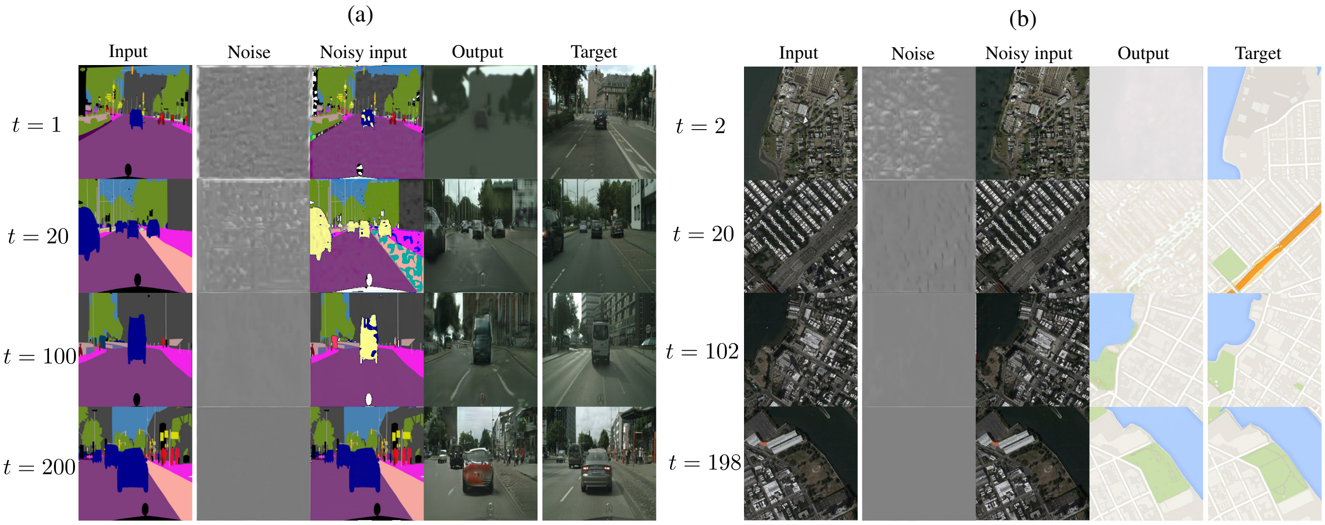

Figure 1. ATME is a GAN where, for each iteration at epoch t, the input images for the generator are corrupted with a noisy representation of the discriminator's decision map at epoch $t - 1$. The adversarial game enforces this noise to be removed, sending the proper signals for equilibration by encouraging the discriminator to converge towards its maximum entropy state. Convergence is often slower in the cases where the noise significantly affects the semantic content of the input (a) and is faster in the opposite cases (b).

Abstract

Generative adversarial networks (GANs) are successfully used for image synthesis but are known to face instability during training. In contrast, probabilistic diffusion models (DMs) are stable and generate high-quality images, at the cost of an expensive sampling procedure. In this paper, we introduce a simple method to allow GANs to stably converge to their theoretical optimum, while bringing in the denoising machinery from DMs. These models are combined into a simpler model (ATME) that only requires a forward pass during inference, making predictions cheaper and more accurate than DMs and popular GANs. ATME breaks an information asymmetry existing in most GAN models in which the discriminator has spatial knowledge of where the generator is failing. To restore the information symmetry, the generator is endowed with knowledge of the entropic state of the discriminator, which is leveraged to allow the adversarial game to converge towards equilibrium. We demonstrate the power of our method in several image-to-image translation tasks, showing superior performance than state-of-the-art methods at a lesser cost. Code is available at https://github.com/DLR-MI/atme.

1. Introduction

Recent advances in deep learning have led to remarkable progress in the field of image synthesis. Among the most exciting applications, image-to-image translation (where an image in domain A is transformed into a different domain B while preserving the original semantic content) has played a prominent role [30]. This task is often addressed using GANs [12] or, more recently, with DMs [14]. Although DMs have been shown to produce high-quality images with unprecedented success, it does so after sequential sampling over multiple time steps. On the other hand, GANs require only a single forward pass for prediction, but suffer from training instabilities that hinder performance.

In this paper, we propose a novel model for image-to-image translation that harnesses the high-quality image generation power of DMs while eliminating their time-sampling limitation using a GAN. Our approach recognizes that the training instabilities in the latter are rooted in a phenomenon similar to the violation of the second law of thermodynamics by Maxwell's demon [26, 31], and suggests a simple solution to avoid this.

In order to achieve stable training, we build a GAN whose generator receives images corrupted by a noisy representation of the discriminator's decision map — as it traverses the training epochs, but not across an independent time-axis as in DMs. The generator then learns to denoise its input to produce the output image, enforcing the discriminator's convergence to its maximum entropy state, as shown in practice by the approach of the GAN (on average) to its theoretical optimum corresponding to the Nash equilibrium [9].

By learning to diffusively attend to the discriminator mean entropy, our model (ATME) helps to improve training stability by breaking the information asymmetry between the generator and discriminator, leading to better performance in image-to-image translation tasks.

The main contributions of this work are therefore:

- A novel model that fuses the sampling strengths of GANs with the core denoising ideas of DMs into a single efficient model for image-to-image translation.
- A practical and simple measure of convergence of GAN models, consistent with the original theoretical description of their optimality.

Our approach builds on recent advancements in the field, particularly from diffusion models. These have achieved state-of-the-art performance in image generation [20]. Nevertheless, they require thousands of model evaluations to generate high-quality samples [5, 33, 34]. Bridging the gap with GANs is therefore an important step towards enabling high-quality and efficient image-to-image translation for a range of practical applications.

2. Related work

GANs for image-to-image translation. GANs have been for a long time the de facto method for generation of synthetic images [30]. pix2pix [17] was the first unified framework for supervised image-to-image translation using conditional GANs. It serves as a foundational model on top of which other solutions have been built, such as adding cycle consistency to a couple of GANs, *i.e.* CycleGAN [42], for unsupervised image-to-image translation. The latter has further inspired other models such as UNIT [22], which leverages a latent representation of the support of the joint distribution of the unpaired images, and several other multi-domain variants [6, 15, 23]. Of special importance for this work is the use of attention mechanisms in GANs. In particular, FAL [16] improves image synthesis with a generator that repeatedly receives feedback—in several forward passes—from the discriminator. SPA-GAN [8], computes attention in the discriminator to help the generator focus on the most discriminative regions between source and target domains. Most recently, ASGIT [21] also enforces spatial guidance by adding attention in the discriminator, surpassing previous methods for supervised and unsupervised image-to-image translation.

Our approach builds on pix2pix, recognizing the information advantages of its patch discriminator, which is counterbalanced by adding attention to the generator. Since the main focus in this work is the effect on convergence, we study this in a supervised setting.

Convergence during GAN training. Several proposals have been made to address the stability issues posed by training GANs, which include vanishing or exploiting gradients and mode collapse. These typically manifest as an ill-behaved Jacobian of the gradient vector field of the associated GAN objectives [28]. To address this, SNGAN [29] proposes a weight normalization technique called spectral normalization to stabilize the training of the discriminator. On the other hand, WGAN [1] introduces the Wasserstein distance between real and fake distributions as an objective to optimize, alleviating the mode collapse problem of vanilla GANs [12], which optimize the Jensen-Shannon divergence. WGAN-GP [13] improves training in practice by adding a gradient penalty to enforce the required discriminator 1-Lipschitz constraint. Alternatively, LSGAN [25] optimizes the Pearson χ^2 divergence between the real and fake distributions. Viewing the convergence in GAN training as a matter of finding the right divergence to minimize at each step is misleading though [10]; more beneficial convergence characteristics are found in practice by adding instance noise or gradient penalties [27].

Architectures also play a role in the stability of GAN training. Energy-based GANs view the discriminator as an energy function taking on lower values for regions near the data manifold. By using the reconstruction error of an autoencoder as an energy function, EBGANs [41] exhibit more stable behavior than vanilla GANs. After approximating the Wasserstein distance using autoencoders, BEGAN [2] intends to balance the generator and discriminator during training. RGANs [18] make the discriminator relativistic (i.e. discriminating whether real data is more realistic than fake data) making training more stable.

Rather than improving network architectures, or changing the objectives functions for training, or regularizing gradients/weights; our work focuses on vanilla GANs with standard networks, stabilizing training by symmetrizing the information exchange between the GAN adversaries.

Diffusion probabilistic models. Diffusion models are generative models that iteratively transform a random noise distribution into a target data distribution by learning a reverse denoising process [14, 35, 36]. They have arisen as the current state of the art in the field of synthetic data generation [40], surpassing GANs [20] in the quality of image synthesis — after denoising either directly in the image space [19] or in a latent representation, such as latent diffusion models (LDMs) [32]. In the context of high-quality image generation leveraging intance noise injection, combining ideas from diffusion models with GANs has gained current research traction [38, 39]. However, the cost of the sampling procedure in diffusion-based models still remains an issue, which may be mitigated by modeling the denoising distribution as a complex multimodal distribution instead of a Gaussian [39], or by making the number of timesteps dependent on the data and the generator [38].

Our approach for injecting instance noise is not based on an independent and expensive diffusion process. It is rather the iterative visit of the data distribution through the training epochs that occurs diffusively, after corrupting the generator inputs with a representation of the disorder state of the discriminator outputs.

3. Background

3.1. Conditional GANs

The generator G in these models learn a mapping from the image x and noise vector z to the image y. Its output is discriminated by a model D, judging whether the image is real or fake. The objective is

$$\begin{aligned}\tilde{\mathcal{L}}_{GAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] +& \\ \mathbb{E}_{x,z}[\log(1-D(x,G(x,z)))],&\end{aligned} \tag{1}$$

where G is trained to minimize this objective and D is trained to maximize it, known as the min-max game.

With the introduction of the patchGAN discriminator in pix2pix [17], the discriminator outputs a tensor (default size of 30×30), with each entry D_i classifying a patch (receptive field size of 70 × 70) in the input image. With N being the number of patches, the objective becomes

$$\mathcal{L}_{GAN}(G,D) = \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathcal{L}}_{GAN}(G,D_i). \tag{2}$$

The motivation of discriminating by patches is enforcing the generator to produce correct high-frequency patterns, while the low frequencies are captured by a $L1$ penalty

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|G(x,z)-y\|_1]. \tag{3}$$

The final objective is then

$$\arg\min_G \max_D \ \mathcal{L}_{GAN}(G,D) + \lambda\,\mathcal{L}_{L1}(G), \tag{4}$$

with λ typically chosen as $\lambda = 100$.

3.2. Diffusion models

Diffusion models are generative models designed to learn a data distribution $p(y_0)$ by sequentially denosing a normally-distributed variable $y_t \sim \alpha_t y_0 + \sigma_t \varepsilon$, by using a model $y_\theta = y_\theta(y_t, t)$ with the objective

$$\mathcal{L}_{DM} \propto \mathbb{E}_{\varepsilon,t}\|y_0 - y_\theta(y_t,t)\|^2. \tag{5}$$

Here the sequences $(\alpha_t)_{t=1}^T$ and $(\sigma_t)_{t=1}^T$ are chosen following a schedule that makes the signal-to-noise ratio $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ small enough at $t = T$ (typically $\text{SNR}(T)\,\|y_0\|^2 \sim 10^{-5}$) and $\varepsilon \sim \mathcal{N}(0,1)$. In practice, only one schedule for the variable β_t in $\bar{\alpha}_t = \prod_{s=1}^t (1-\beta_s)$ is chosen, with $\text{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$.

4. Attending the mean entropy (ATME)

Introducing a discrimination by patches allows the discriminator to have a notion of *where* the generator is failing. This makes the min-max game asymmetric in favor of the discriminator, since the generator has no direct spatial clue of where the discriminator is failing. Without further intervention, this forbids a proper equilibration of the game, resulting in a lack of convergence. Our task is to find the piece of information about the discriminator that the generator should know in order to recover the symmetry.

The situation is similar to the information asymmetry introduced in statistical physics by Maxwell's demon. That is, when two ideal gases at different temperatures are placed in separate containers communicated by a switchable hole, equilibration (corresponding to the maximum entropy state) is achieved when the hole is opened — more fast-moving particles moving from the hot container to the cold one than backwards. But if an entity (demon) is introduced, which opens the hole to allow the backward motion and close it to block the forward, the cold container will be colder and the hot container hotter, and equilibration never takes place.

In the GAN game, the information gain introduced by the patch discriminator is analog to the information gain of Maxwell's demon due to its knowledge of the velocity of the particles in both containers. We propose to incentivate a proper equilibration by letting the generator enforce the corresponding maximum entropy state — seeing the Nash equilibrium [9] as a thermal equilibrium. The following fact (proved in the appendix) hints us on how to achieve this:

Theorem 1. *Let Y_i be a binary random variables taking on the value $y_i = 1$ with probability D_i. If they are statistical independent, the joint distribution $P(Y_1, \cdots, Y_N)$ has maximum entropy if and only if $D_i = \frac{1}{2}$ for all i. In this state, the objective in Eq. (2) reaches the value $-\log(4)$ for an optimal discriminator and generator.*

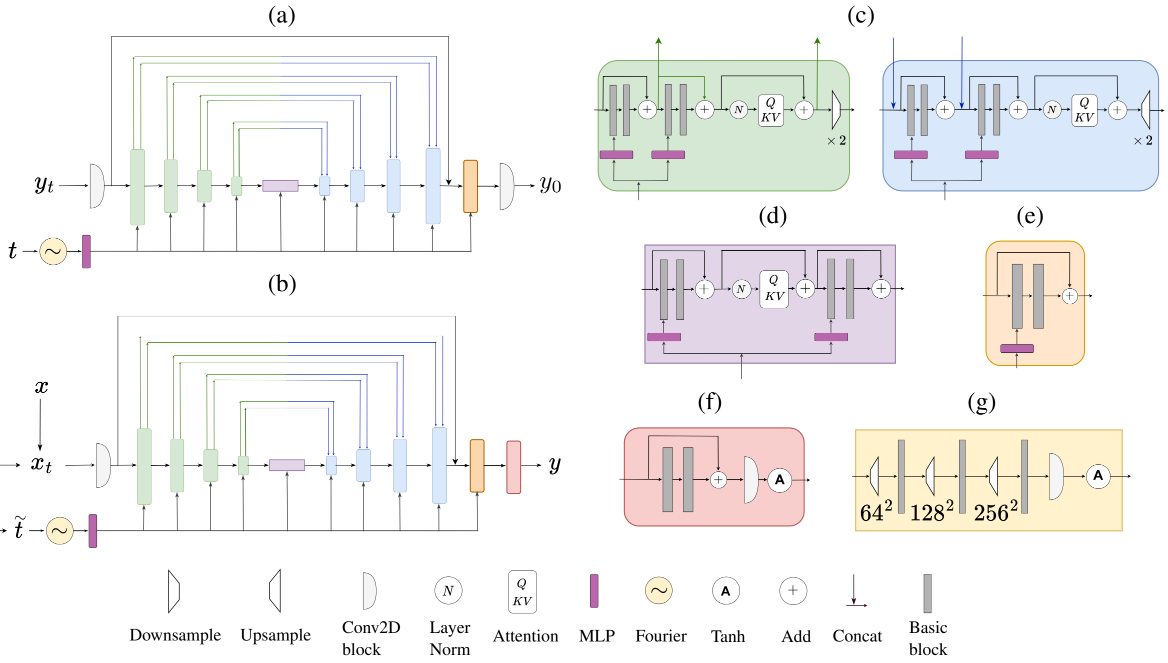

Figure 2. Building blocks (c)-(e) for the denoising UNet (a) used in diffusion models [14]. Our model (b) adds new building blocks for its generator, consisting of (g) for listening to the discriminator entropic state in order to corrupt the input image with noise, and a modified head (f) to remove spurious high-frequency patterns. All building blocks are based on a basic block that convolves the input, normalizes the resulting channels, optionally applies an affine transformation (using the Fourier features), and finally applies SiLU activation.

We propose to endow the generator with a notion of the state of "disorder" of the discriminator decisions, this being a surrogate to its entropy. Denoting by D_t the output tensor of the discriminator at training epoch t (having entries $D_{i;t}$), we introduce the learnable mapping $W_t = W(D_{t-1})$ having a range in the space of the input images of the generator. This should have the following properties:

- As D_t tends to the maximum entropy state, $W(D_t)$ tends to a constant tensor, and viceversa. That is, as $D_{i;t} \to \frac{1}{2}$ for all patches i, $W_{r;t} \to w$ for all pixels r. This is our statement of the preservation of the state of disorder under the action of W.
- The differences $W(D_t) - W(D_{t-1})$ are uncorrelated in time and approximately Gaussian.

The second property is a weaker one, only ensuring that the input images for the generator, which we take as

$$x_t = x_0 + x_0\, W(D_{t-1}), \tag{6}$$

initially follow a Brownian motion, diffusing through the epochs during training. This allows us to borrow the intuition from the diffusion models. That is, we corrupt the input image $x_0 = x$ with "noise" arising from $W(D_{t-1})$ and train the generator to get rid of this noise in order to capture the correct mapping $x \to y$ (as shown in Fig. 1). As a side effect, removing this noise sends the signal to the discriminator to seek the maximum entropy state, by the main property of $W(D_t)$.

It is important to note that, although the epochs index the time steps t in the experiments, the arrow of time set in the generator has to follow the discriminator's entropic state. This is achieved by estimating the temporal position of the noising events according to

$$\tilde{t} = \mathbb{E}\,[W(D_{t-1})], \tag{7}$$

which is similar to the ordering imposed in the diffusion models by SNR(T) being small and SNR(0) being large.

The loss of ATME at epoch t is then, similar to Eq. (4),

$$\mathcal{L}^t_{\text{ATME}}(G, D) = \mathcal{L}^t_{GAN}(G, D) + \lambda\, \mathcal{L}^t_{L1}(G), \tag{8}$$

with the superindex indicating that the variables (x, z) are replaced by the combined variable x_t in Eq. (6), and the generator acquires the functional form (see Eq. (5)) that is used in the diffusion models, $G(x_t) := y_\theta(x_t, \tilde{t})$.

At inference, D_{t-1} is sampled element-wise from a normal distribution with mean $\frac{1}{2}$ (the maximum-entropy value) and small standard deviation (set to 0.001 in all experiments).

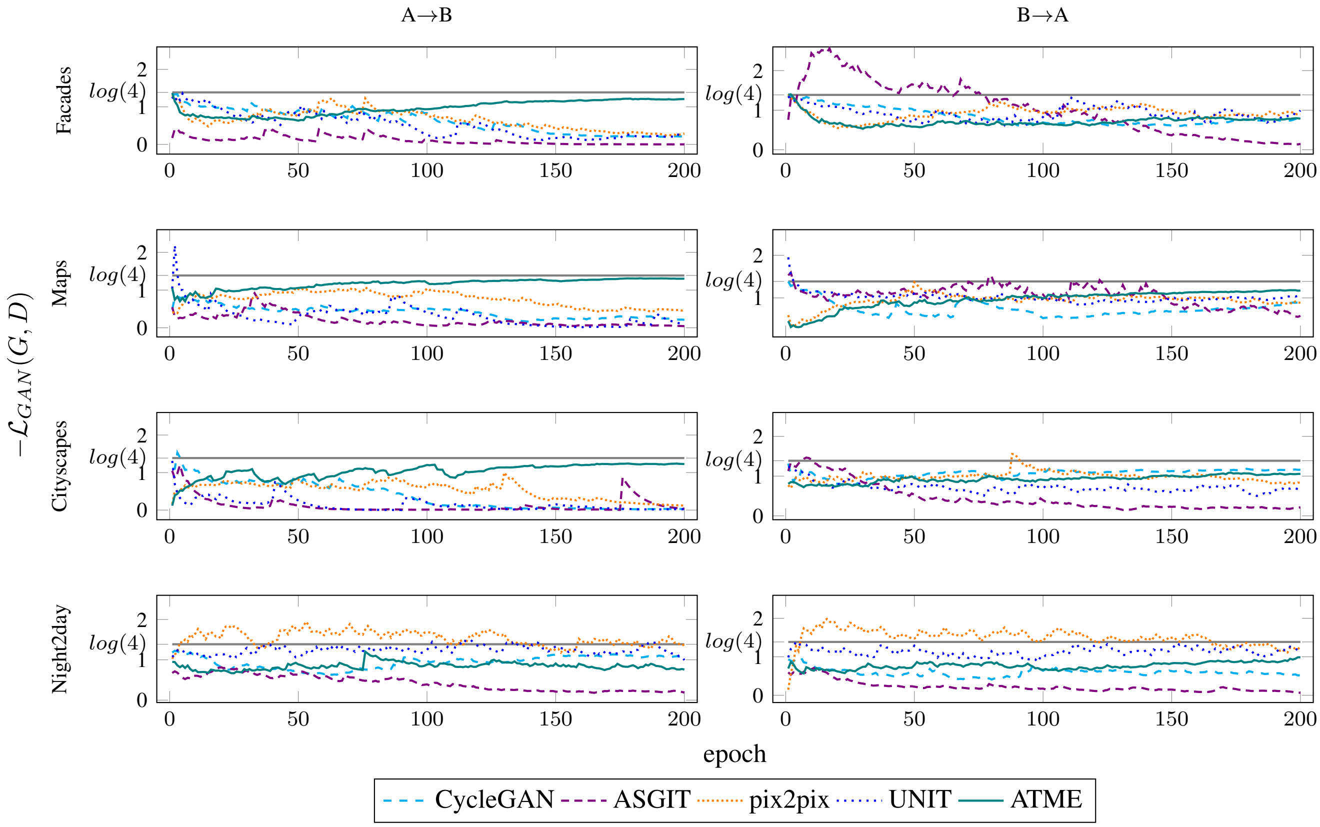

Figure 3. Smoothed $-\mathcal{L}_{GAN}(G, D)$ from Eq. (2) at the end of each training epoch for all GAN models and datasets considered. According to Eq. (4) and Theorem 1, this should converge — in the limit of a large enough model and infinite data [11] — to the Nash equilibrium, where the value $\log(4)$ is reached. Being the lightest of all, our model ATME shows better converge properties (on average) except in the largest dataset Night2day, where it lacks capacity to accommodate all the data variability. Also, its convergence is slower in the tasks B→A where it is harder to remove the noise applied to the input images due to this noise significantly altering their semantic content.

4.1. Model architecture

The architecture of the patch discriminator in ATME follows the implementation of pix2pix [17]. The generator is shown in Fig. 2(b). It has the UNet structure used in the diffusion models (see Fig. 2(a)) with additional blocks that we introduce to listen to the discriminator's entropic state, Fig. 2(g), and a modified head (Fig. 2(f)) to remove spurious high-frequency patterns.

The UNet is mainly parameterized by an embedding dimension d and attention resolutions $R = (r_1, r_2, \cdots, r_H)$, with H being half the depth of the network (excluding the middle block). The notation means that at the ith downsampling layer, the number of feature maps go from $d\,r_{i-1}$ to $d\,r_i$ (with $r_0 = 1$). The default network for all the experiments has $d = 64$ with $R = (1, 2, 4, 8)$, corresponding to a network with four downsampling layers, a middle block, and four upsampling layers, as shown in Fig. 2(b).

Dataset	**A**	**B**
Facades	Photo	Architectural labels
Maps	Aerial photo	Map
Cityscapes	Photo	Semantic labels
Night2day	Night photo	Day photo

Table 1. Datasets used in this work. The corresponding images are paired, *i.e.* the first half of the width of the image is called A and the second half is called B.

5. Experiments

5.1. Datasets

We use four of the standard datasets for supervised image-to-image translation: Facades, Maps, Cityscapes, and Night2day, whose details can be found in [17]. For Night2day, we train only on 5000 images. Image-to-image translation is performed in both directions A→B and B→A, as defined in Table 1.

Model	**# Params**	**Facades**		**Maps**		**Cityscapes**		**Night2day**	
	[M]	A→B	B→A	A→B	B→A	A→B	B→A	A→B	B→A
pix2pix	57	31.3 ± 2.3	11.0 ± 0.7	25.7 ± 2.0	19.0 ± 1.8	16.0 ± 0.8	7.8 ± 1.0	19.2 ± 1.6	11.6 ± 1.1
CycleGAN	114	28.1 ± 2.1	18.2 ± 1.0	60.5 ± 1.3	10.8 ± 1.4	45.0 ± 1.0	16.9 ± 1.3	**12.6 ± 2.0**	**9.0 ± 0.9**
UNIT	116	47.9 ± 2.4	18.0 ± 0.9	30.1 ± 0.9	9.3 ± 1.1	16.6 ± 1.1	12.8 ± 1.0	15.6 ± 2.1	19.7 ± 1.5
ASGIT	57	22.6 ± 1.6	**4.9 ± 0.8**	9.2 ± 1.2	7.7 ± 1.2	16.0 ± 1.3	**4.2 ± 0.5**	17.5 ± 2.2	11.1 ± 1.3
LDM	270	30.9 ± 2.4	23.0 ± 1.0	7.9 ± 1.0	10.3 ± 1.3	**5.6 ± 0.6**	5.6 ± 0.5	19.6 ± 2.2	11.9 ± 1.3
ATME	**39**	**18.4 ± 1.8**	9.4 ± 0.7	**2.8 ± 0.6**	**2.8 ± 0.7**	6.5 ± 1.0	5.7 ± 0.8	19.7 ± 2.1	18.3 ± 1.4

Table 2. KID scores (scaled by 100) for the methods evaluated in the datasets shown (lower is better). The best result per column is shown in bold. Our model ATME, shows superior performance, defined as the number of times it has the best KID per task.

Model	**Per-pixel acc.**	**Per-class acc.**	**Class IoU**
pix2pix	0.63	0.18	0.13
CycleGAN	0.49	0.13	0.09
UNIT	0.48	0.12	0.09
ASGIT	0.54	0.17	0.11
LDM	0.57	0.17	0.11
ATME	**0.64**	**0.19**	**0.14**
Ground truth	0.80	0.26	0.21

Table 3. FCN scores (higher is better) after training on Cityscapes B→A at a resolution of 256×256. The best result per column is shown in bold.

5.2. Baselines

Since our model is built using the pix2pix framework, we train the latter for comparison. Additionally, we train CycleGAN [42] (despite its introduction for unsupervised problems) as a reference of a generator that is trained to have an approximate inverse mapping. The hypothesis is that adding cycle consistency may improve convergence since this restricts the possible paths to equilibrium, with respect to those allowed by the highly under-unconstrained source-to-target mapping originally present in pix2pix. Finally, we train the supervised version of ASGIT [21] as a reference of a state-of-the-art model using attention in the discriminator, as well as their implementation of UNIT (at a 256×256 resolution) using a 2-branch residual attention network [37] in the discriminator.

On the other hand, due to the diffusion in a latent-space representation of the target images being more efficient than in the image space, we choose to train an LDM [32] conditioned on the source images for comparison.

5.3. Training details

We train all GAN models from scratch using the default configuration in pix2pix. That is, we use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, with an initial learning rate of 0.0002 for both the generator (UNet-256) and discriminator (patchGAN) of the vanilla GANs. The learning rate is kept constant in the first 100 epochs and linearly decayed to zero for the following 100 epochs. A batch size of 48 is used and instance normalization. Random jittering and horizontal flipping is applied during training to the images with resolution 256×256. For ATME, the UNet-256 is replaced by the UNet in Fig. 2(b) with an embedding dimension of $d = 64$ and resolutions $R = (1, 2, 4, 8)$.

On the other hand, we train the LDMs with the default configuration in [32] for the input resolution 256×256. That is, the denoising is done by the UNet in Fig. 2(a) after downsampling the input (target) images by a factor of $f = 4$ (using the VQ-reg encoder with attention) running the diffusion process, and concatenating the output of this process with a spatially-scaled version of the conditioning (source) image. The diffusion follows a linear schedule of β_t, from $\beta_1 = 0.0015$ to $\beta_T = 0.0205$ in $T = 1000$ timesteps.

5.4. Metrics

We follow recent practices [4, 21] and report the Kernel Inception Distance (KID) between feature representations of real and fake images. The feature extraction [3] is done by the Inception v3 model. Additionally, the FCN score [17] is computed to further evaluate details of the performance in the Cityscapes dataset. This measures the accuracy of the FCN-8s semantic classifier [24] (pre-trained on real images) after segmenting the generated images and comparing the results against the labels these images were synthesized from.

5.5. Evaluating convergence of GANs

We keep track of the loss in Eq. (2) at the end of each epoch and notice that, by Theorem 1 and Eq. (4), the convergence to equilibrium is manifested as the approach of $-\mathcal{L}_{GAN}(G, D)$ to $\log(4)$ during optimization. This is shown in Fig. 3, where ATME shows stable convergence in most cases. The cases where convergence seems slower is presumably due to ATME not being large enough — since GANs are designed to reach Nash equilibrium with a large

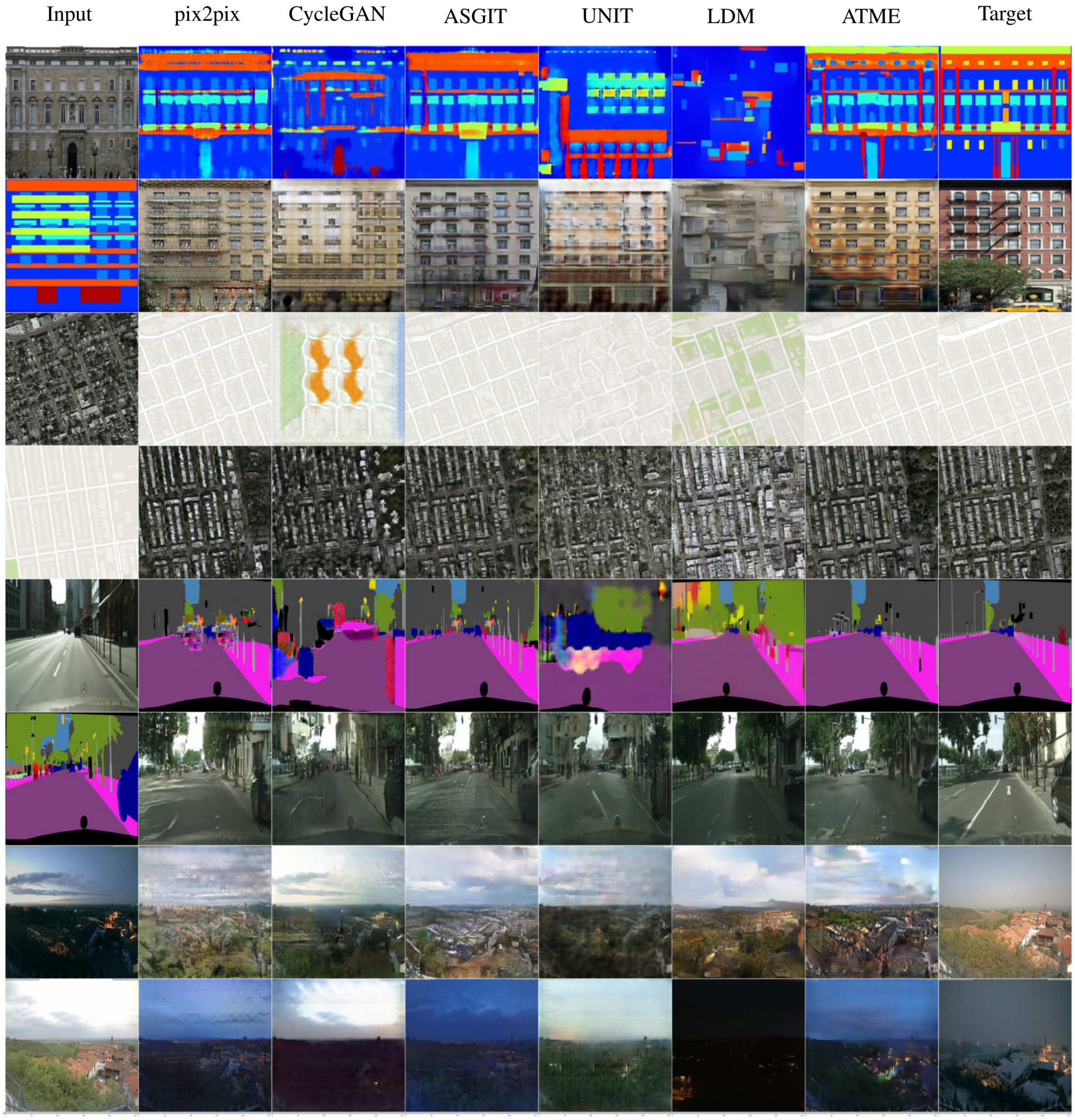

Figure 4. Sample predictions of all models in the evaluated datasets.

enough model and infinite data [11] — or being harder for the generator to remove the input noise in the required number of epochs. The latter is evident from the success in the convergence for the column A→B in Fig. 3, which represents the corruption with noise of the photo (much more semantic content than the labels) according to Table 1. The exception is the Night2day dataset, for which the photo with more semantic content is in the opposite side B. This slowness in noise removal is further illustrated in Fig. 1 where, at the same epoch $t = 20$, the noise in Cityscapes B→A still has much more structure than the noise in Maps A→B.

As mentioned above, the model capacity may also play a role, specially for the largest dataset. As seen in Table 2, ATME is the lightest model so it may not be able to accomodate all the variability of the data distribution in this case. We plan to investigate this further in the future. However, preliminary results show that by enlarging ATME to the configuration $d = 64$ and $R = (1, 1, 2, 2, 4, 4, 4, 8)$, which brings the model to a capacity similar to pix2pix (*i.e.* with ∼57M parameters), the worst KID in Table 2, obtained in the task Night2day A→B, is lowered to 16.3 ± 2.1, taking ATME from the last place to the top-3 after the bigger CycleGAN and UNIT models. These bigger models were observed to suffer mode collapse for some tasks, as evidenced in Fig. 4.

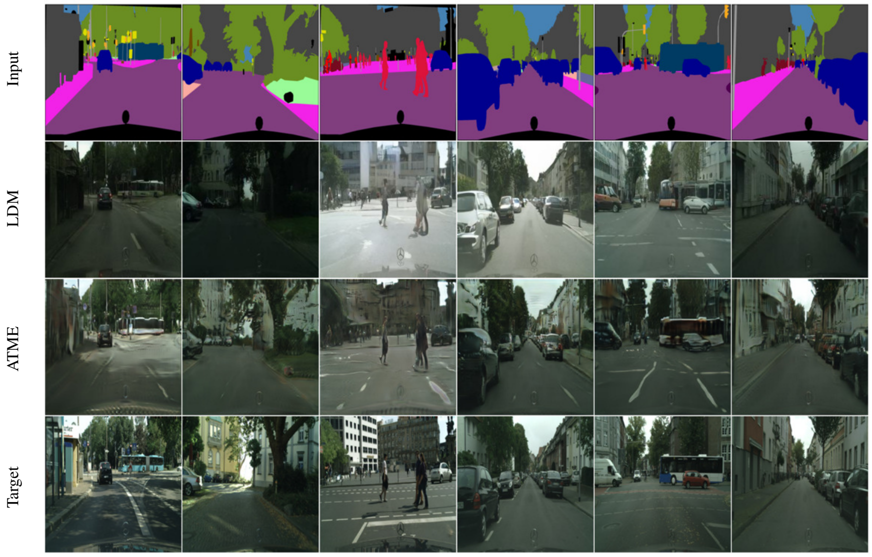

Figure 5. Qualitative predictions of LDM and ATME for a subset of test images in the Cityscapes dataset.

5.6. Quality of image synthesis

Table 2 shows the KID scores for all models and datasets. Despite being the lightest model, our model ATME shows superior performance than the other methods, assessed as the number of times that it has the lowest KID per task.

The quality of image generation is further evaluated using the FCN scores in the Cityscapes dataset (see Table 3), confirming the superiority of ATME compared to the other methods. Sample predictions from all methods in all datasets are shown in Fig. 4.

5.6.1 Distribution Modes: GANs vs Diffusion models

Both GANs and diffusion models are trained to learn the target distribution conditioned on the source images. Given an input image x, the models are expected to output the most probable image $\hat{y}$ sharing semantic content with x. This should have a strong similarity with the ground truth y. Although the diffusion models are known to predict images with very high quality, surprisingly for us, the predictions are far from the right mode, as can be seen in Fig. 5, *i.e.* LDM not being able to understand the semantics of the right pose (*e.g.* car facing inwards being confused with the car facing outwards), the right contrast, etc. This explains the results of Table 3 and suggests that GAN models are more appropriate for supervised image-to-image translation than diffusion models.

6. Conclusion

We have shown that a significant improvement in the convergence properties of GANs for image-to-image translation is achieved when making the generator and discriminator exhange information symmetrically. We achieve this by informing the generator about the entropic state of the discriminator, as a guide for the equilibration of the adversarial game. The quality of image synthesis is high compared to state-of-the-art methods and our model ATME predicts the modes of the conditional target distribution better than diffusion models.

Several research directions are left open, including exploring a generator model in ATME with higher capacity and, most importantly, extending the method to unsupervised image-to-image translation.

Appendix

To avoid clutter in notation, we omit the condition on x in the following.

Theorem 1. *Let Y_i be binary random variables taking on the value $y_i = 1$ with probability D_i. If they are statistical independent, the joint distribution $P(Y_1, \cdots, Y_N)$ has maximum entropy if and only if $D_i = \frac{1}{2}$ for all i. In this state, the objective in Eq.* (2) *reaches the value $-\log(4)$ for an optimal discriminator and generator.*

Proof. The joint entropy becomes the sum of the marginal entropies $-\sum_i \sum_{y_i} D_i(y_i) \log D_i(y_i)$ if and only if the random variables $Y_1, \cdots, Y_N$ are statistically independent [7], which is implicit in the patch discriminator being Markovian [17]. Now, the entropy of a binary random variable is known to reach a maximum when $D_i(y_i) = \frac{1}{2}$ for all i and y_i. In this case, the objective in Eq. (2) collapses to

$$\mathcal{L}_{GAN} = \frac{1}{N} N \left(\log\left(\tfrac{1}{2}\right) + \log\left(\tfrac{1}{2}\right) \right) = -\log(4), \qquad (9)$$

corresponding to the value for an optimal discriminator and generator [12]. □

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 214–223, 2017. 2

[2] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 2

[3] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 6

[4] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 6

[5] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 2

[6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[7] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, N.J, second edition edition, 2006. 8

[8] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia*, 23:391–401, 2021. 2

[9] Farzan Farnia and Asuman Ozdaglar. Do GANs always have Nash equilibria? In *Proceedings of the 37th International Conference on Machine Learning*, pages 3029–3039, 2020. 2, 3

[10] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. In *International Conference on Learning Representations*, 2018. 2

[11] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 5, 7

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1, 2, 8

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, 2017. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1, 3, 4

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[16] Minyoung Huh, Shao-Hua Sun, and Ning Zhang. Feedback adversarial learning: Spatial feedback for improving generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 5, 6, 8

[18] A. Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard gan. In *International Conference on Learning Representations*, 2019. 2

[19] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

[20] Jaakko Lehtinen, Marko Järvenpää, Niels Kasenburg, Antti Honkela, and Mathias Berglund. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 3

[21] Yu Lin, Yigong Wang, Yifan Li, Yang Gao, Zhuoyi Wang, and Latifur Khan. Attention-based spatial guidance for image-to-image translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 816–825, January 2021. 2, 6

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 2

[23] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 6

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 2

[26] Koji Maruyama, Franco Nori, and Vlatko Vedral. Colloquium: The physics of maxwell's demon and information. *Rev. Mod. Phys.*, 81:1–23, Jan 2009. 2

[27] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 2

[28] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett,

editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[29] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2

[30] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2022. 1, 2

[31] Andrew Rex. Maxwell's demon—a historical review. *Entropy*, 19(6), 2017. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 3, 6

[33] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 2

[34] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[35] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 3

[36] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2019. 3

[37] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 6

[38] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion, 2022. 3

[39] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans, 2021. 3

[40] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2022. 3

[41] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *International Conference on Learning Representations*, 2017. 2

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2, 6