UVCGAN v2: An Improved Cycle-Consistent GAN for Unpaired Image-to-Image Translation

Dmitrii Torbunov, Yi Huang, Huan-Hsin Tseng, Haiwang Yu,
Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, Yihui Ren
Brookhaven National Laboratory, Upton, NY, USA
{dtorbunov,yhuang2,htseng,hyu,jhuang,sjyoo,mlin,bviren,yren}@bnl.gov

Abstract

An unpaired image-to-image (I2I) translation technique seeks to find a mapping between two domains of data in a fully unsupervised manner. While initial solutions to the I2I problem were provided by generative adversarial neural networks (GANs), diffusion models (DMs) currently hold the state-of-the-art status on the I2I translation benchmarks in terms of Fréchet inception distance (FID). Yet, DMs suffer from limitations, such as not using data from the source domain during the training or maintaining consistency of the source and translated images only via simple pixel-wise errors. This work improves a recent UVC-GAN model and equips it with modern advancements in model architectures and training procedures. The resulting revised model significantly outperforms other advanced GAN- and DM-based competitors on a variety of benchmarks. In the case of Male-to-Female translation of CelebA, the model achieves more than 40% improvement in FID score compared to the state-of-the-art results. This work also demonstrates the ineffectiveness of the pixel-wise I2I translation faithfulness metrics and suggests their revision. The code and trained models are available at https://github.com/LS4GAN/uvcgan2.

1. Introduction

Image-to-image (I2I) translation models aim to find a mapping between two domains of images. When paired examples of images from two domains are available, such a mapping can be easily solved in a supervised manner. However, the unpaired I2I translation, where examples of pairs are not available, poses a more interesting problem. The ability to perform an unpaired I2I translation is highly beneficial as obtaining paired datasets in the real world is often difficult, time-consuming, or even impossible [4].

The advancement of unpaired I2I largely benefits from recent developments in deep generating models, such as (variational) Autoencoder [18, 25], generative adversarial networks (GANs), and generating flows [12, 13, 24, 37]. One early successful unpaired I2I model is CycleGAN [52], which uses a cycle-consistency constraint that requires a cyclic back-and-forth translation between two domains to produce the original image. Several succeeding models inspired by CycleGAN, such as STARGAN [7, 8], SEAN [53], U-GAT-IT [23], and CUT [34], are designed to further enhance the quality and diversity of the generated images. However, GAN-based I2I methods lag behind general developments in GAN architecture and training procedures [21, 40].

Meanwhile, *diffusion models* (DMs) provide an alternative route to image generation [19]. With a recent spike of interest in such models, several applications of DMs to unpaired I2I translation have been developed [6, 30, 48]. Despite being relatively new, the DM-based EGSDE [48] approach already has demonstrated superior results on several benchmarks. However, because they do not use source images during the training, DM-based solutions may perform a suboptimal translation [48]. Additionally, DM-based methods rely on pixel-wise L_2 distances to maintain the consistency of the source and translated images. This simple consistency measure is not guaranteed to preserve any semantically meaningful features and can restrict image transformations.

On multiple occasions, revisiting a classic neural architecture and improving it with a number of modern additions has led to vast improvements in performance [2, 21, 28]. Based on this observation, UVCGAN [42] modernizes the classic CycleGAN architecture and achieves state-of-the-art performance on several I2I tasks. Unlike the DM-based models, CycleGAN's training procedure is able to utilize images from the source and target domains effectively and simultaneously. Moreover, CycleGAN maintains an intrinsic consistency between the source and translated images (via the cycle-consistency constraint)—a feature that cannot be achieved by simple pixel-wise consistency measures.

Following the UVCGAN's spirit, we introduce addi-

tional architectural innovations and bring modern training techniques to CycleGAN. The resulting model significantly outperforms both UVCGAN (e.g., 50% better realism on CelebA male-to-female) and the state-of-the-art DM-based EGSDE (e.g., 40% better on CelebA-HQ male-to-female). The following summarizes the contributions of this work:

- We introduce a novel hybrid neural architecture for image translation, combining benefits of ViT and Style Modulated Convolutional blocks.
- We propose enhancing a common I2I discriminator architecture with a specialized head to prevent the problem of mode collapse [15].
- We further improve the UVCGAN's training procedure by incorporating modern GAN training techniques that are not commonly applied to unpaired I2I problems.
- We perform extensive quantitative evaluations of our model and demonstrate it surpasses state-of-the-art competitors in terms of image realism on various translation tasks.
- Finally, we highlight inconsistencies and limitations of current unpaired I2I evaluation metrics and procedures. In response, we propose improved faithfulness measures that are better aligned with human perception.

2. Related Works

Problems related to unpaired I2I translation have been approached from multiple perspectives. There are two major classes of solutions: GAN- and diffusion-based methods.

GAN-based Methods. Multiple GAN-based methods have been developed to tackle the problem of unpaired I2I translation. One distinct group of GAN-based methods involves methods that rely on cycle consistency, including CycleGAN [52], DualGAN [45], U-GAT-IT [23], and the recent UVCGAN [42]. This class of algorithms requires two generator networks that translate images in opposite directions. Its basis is a cycle-consistency constraint, requiring that a cyclically translated image should match the original. Cycle-consistent models can show remarkable performance [42], but there are concerns that the cycle-consistency condition may be too restrictive.

ACLGAN [49] attempts to relax the cycle-consistency constraint and replace it with a weaker adversarial one. Such relaxation allows the network to make larger changes to the source image, potentially achieving better translation quality. CouncilGAN [32] moves a step further and completely discards cycle consistency. Instead, it trains an ensemble of generators, performing translation in a single direction and allowing for a larger diversity of generated images.

CUT [34] takes an alternate route and uses a contrastive loss to maximize the information between the source and the translated images. This approach removes the need to have multiple generators and allows CUT to train faster. Using CUT as a basis, ITTR [51] improves its performance by modifying the generator architecture. In a similar fashion, LSeSim [50] designs a contrastive-based loss function that guides the image translation without the need for multiple generators.

Diffusion-based Methods. With the recent explosion of interest in DMs, many have attempted to employ them for unpaired I2I translation. For instance, ILVR [6] achieves an unpaired image translation by modifying the standard Gaussian denoising process. It relies on a DM trained only on the target domain but guides it toward the source image during denoising.

SDEdit [30] introduces another viable approach for performing image translation. Instead of modifying the diffusion process itself, it simply changes the starting point of diffusion. SDEdit uses a source image perturbed by Gaussian noise as a seed image and runs the standard diffusion process on top of it.

Finally, EGSDE [48] makes an observation that both ILVR and SDEdit are trained on the target domain data. As such, they may perform a suboptimal translation. EGSDE combines ILVR and SDEdit approaches and modifies both the starting point of the denoising process and the denoising process itself. To overcome the limitation of the DM being trained only on the target domain, it introduces a special energy function, pretrained on both domains. This energy function guides the denoising process, allowing it to achieve state-of-the-art results on several benchmarks.

3. Method

UVCGANv2 revisits the classic CycleGAN [52] architecture. UVCGANv2 inherits several advancements from UVCGAN [42], such as a hybrid U-Net-ViT (vision transformer) generator architecture, self-supervised generator pre-training, and better training strategies. This section describes several improvements that exceed UVCGAN, including the generator, discriminator and the training procedure.

3.1. Review of the Original UVCGAN

UVCGAN follows the CycleGAN framework [23, 52] that interlaces two generator-discriminator pairs for unpaired I2I translation. Denote the two domains by A and B. Generator $\mathcal{G}_{A \to B}$ translates images in A to resemble those from domain B. Discriminator $\mathcal{D}_B$ distinguishes images in B from those translated from A. The same goes for the other translation direction, $\mathcal{G}_{B \to A}$ and $\mathcal{D}_A$. The discriminators are updated by backpropagating loss in distinguishing real and translated (fake) images (called *GAN loss*):

$$\mathcal{L}_A^{\text{disc}} = \mathbb{E}_B \ell_{\text{gan}} \left(\mathcal{D}_A \left(a_{\text{f}} \right), 0 \right) + \mathbb{E}_A \ell_{\text{gan}} \left(\mathcal{D}_A (a), 1 \right), \quad (1)$$

$$\mathcal{L}_B^{\text{disc}} = \mathbb{E}_A \ell_{\text{gan}} \left(\mathcal{D}_B \left(b_{\text{f}} \right), 0 \right) + \mathbb{E}_B \ell_{\text{gan}} \left(\mathcal{D}_B (b), 1 \right) \quad (2)$$
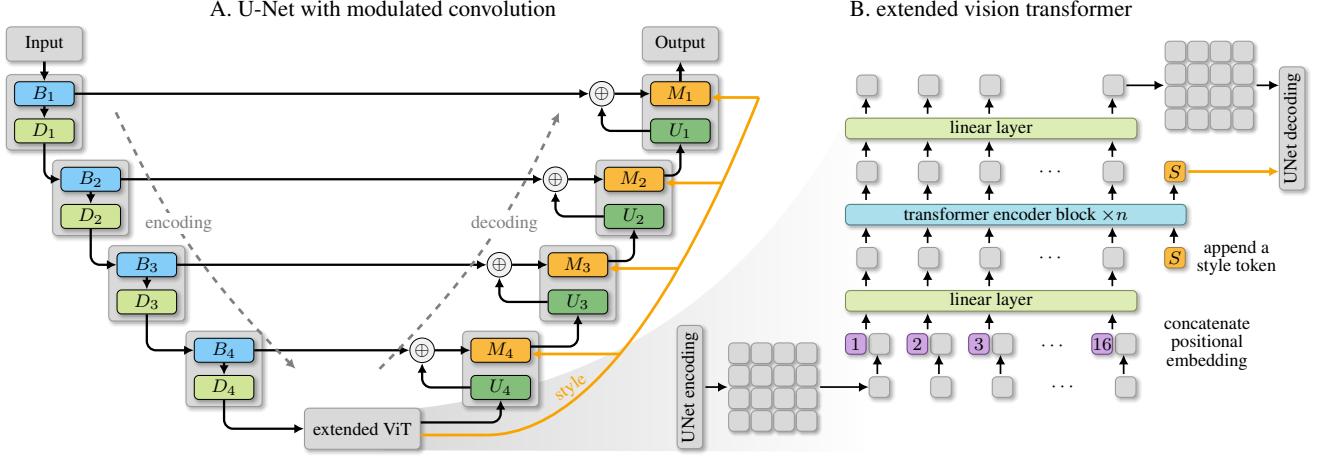
Figure 1. **UVCGANv2 Generator.** UVCGANv2's generator is a U-Net (Panel A) with an extended vision transformer bottleneck (eViT, Panel B). The eViT outputs a style token for the modulated convolution blocks [21] (M_i, $i = 1, 2, 3, 4$) in the decoding path of the U-Net. Refer to [42] for details about the input layer, output layer, basic block B_i, downsampling block D_i and upsampling block U_i in the U-Net, and transformer encoder block in the eViT.

where a is an image from A, b is an image from B, $b_{\mathrm{f}} = \mathcal{G}_{A \to B}(a)$, $a_{\mathrm{f}} = \mathcal{G}_{B \to A}(b)$ (subscript f means fake), 0 is the label for fake images, 1 is the label for real images, and ℓ_{gan} represents a classification loss function (L_2, cross-entropy, Wasserstein [1], etc.). The generators are updated by backpropagating loss from multiple sources: GAN loss for realistic translation, cycle-consistency loss, and optionally identity loss for within-domain translation. Using domain A as an example, we have:

$$\mathcal{L}_A^{\mathrm{gan}} = \mathbb{E}_A \ell_{\mathrm{gan}} \left(\mathcal{D}_B \left(b_{\mathrm{f}} \right), 1 \right), \tag{3}$$

$$\mathcal{L}_A^{\mathrm{cyc}} = \mathbb{E}_A \ell_{\mathrm{reg}} \left(a_{\mathrm{c}}, a \right), \quad \mathcal{L}_A^{\mathrm{idt}} = \mathbb{E}_A \ell_{\mathrm{reg}} \left(a_{\mathrm{i}}, a \right) \tag{4}$$

where $a_{\mathrm{c}} = \mathcal{G}_{B \to A} \circ \mathcal{G}_{A \to B}(a)$, $a_{\mathrm{i}} = \mathcal{G}_{B \to A}(a)$, and ℓ_{reg} is any pixel-wise loss function (L_1 or L_2, etc.). The generator loss is defined as a linear combination:

$$\mathcal{L}^{\mathrm{gen}} = \left(\mathcal{L}_A^{\mathrm{gan}} + \mathcal{L}_B^{\mathrm{gan}} \right) + \lambda^{\mathrm{c}} \left(\mathcal{L}_A^{\mathrm{cyc}} + \mathcal{L}_B^{\mathrm{cyc}} \right) + \lambda^{\mathrm{i}} \left(\mathcal{L}_A^{\mathrm{idt}} + \mathcal{L}_B^{\mathrm{idt}} \right) \tag{5}$$

3.2. Source-driven Style Modulation

UVCGAN uses a hybrid UNet-ViT generator network. We attempt to increase the generator's performance by redesigning its architecture. We extend the generator, enabling it to infer an appropriate target style for each input image. Then, we modulate [21] the generator's decoding branch by the style, significantly increasing its expressiveness.

Specifically, at the bottleneck of the generator, the image is encoded as a sequence of tokens to be fed to the Transformer network. We augment this sequence with an additional learnable style token S. The state of the S token at the output of the Transformer serves as a latent image style. For each convolutional layer of the U-Net's decoding branch, we generate a specific style vector s_i from S by trainable linear transformations.

The style modulation [21] effectively scales weights $w_{i,j,x,y}$ of the convolutional operator by the supplied style vector s_i, yielding modulated weights:

$$w'_{i,j,x,y} = s_i \cdot w_{i,j,x,y} \tag{6}$$

where i, j refer to the input and output feature maps and x, y enumerate the spatial dimensions. To preserve the magnitude of the activations, the scaled weights $w'_{i,j,x,y}$ need to be demodulated. The demodulation operation further renormalizes the convolution weights as follows:

$$w''_{i,j,x,y} = \frac{w'_{i,j,x,y}}{\sqrt{\sum_{i,x,y} \left(w'_{i,j,x,y} \right)^2 + \epsilon}} \tag{7}$$

where ϵ is a small number to prevent numerical instability.

Our approach is different from StyleGANv2, which generates the style vectors s_i from a random prior. We use a learnable S token of the transformer to infer the required target style from the source image itself. The way S is processed is similar to the [class] token of the ViT [14]. However the [class] token mainly is used for the classification task.

3.3. Batch-Statistics-aware Discriminator

Batch statistics offers an effective way to combat mode collapse and promote diversity for GAN models [20, 38]. We seek to improve the UVCGAN discriminator's performance by integrating a minibatch discrimination technique. However, employing traditional minibatch discrimination
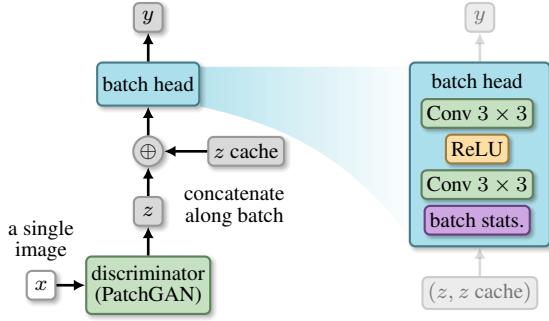
Figure 2. **UVCGANv2 Discriminator with a batch head.** For the batch statistics block, we can use either a standard batch normalization layer or batch standard deviation [20].

requires having a nontrivial batch size. With the recent push to work on high-resolution images, increasing the batch size will strain GPU hardware limitations. We workaround this problem by decoupling the batch size from the minibatch discrimination technique and designing a GPU memory-efficient algorithm.

The main idea is to integrate a cache of past discriminator features to serve as a substitute for a large batch size. Similar ideas, known as "memory bank," have been explored in representation learning models [16, 44]. To our best knowledge, using a memory bank type of cache in a GAN is novel.

We design a composite discriminator made of a main *body* and a *batch head*. The composite discriminator body can be any common discriminator, but, for the purposes of this work, we use UVCGAN's discriminator without the last layer. The batch head is made of a layer that captures batch statistics, followed by two convolutional layers (Figure 2). This modular discriminator architecture allows for easily swapping different discriminator bodies while still preserving the minibatch discrimination power. During training, the batch features are stored in four separate caches: real images from domain a, real images from domain b, and fake images from both domains. All the caches have a fixed size and follow the first-in-first-out (FIFO) update policy.

The discriminator's batch head receives a concatenation (along the batch dimension) of the discriminator body output for the current minibatch along with a history of the past outputs from a cache (Figure 2). We experiment with two types of batch statistics layers: batch standard deviation (BSD) [20] and a simple batch normalization (BN).

3.4. Pixel-wise Consistency Loss

To improve the consistency of the generated and source images, we experiment with the addition of an extra term $\mathcal{L}_{consist}$ to denote generator loss (5). This term captures the L_1 difference between the downsized versions of the source

and translated images. For example, for images of domain A

$$\mathcal{L}_{consist,A} = \mathbb{E}_A \ell_1 \left(F(\mathcal{G}_{A \to B}(a)), F(a) \right) \tag{8}$$

where F is a resizing operator down to 32×32 pixels (low-pass filter). We add this term to the generator loss (5) with a magnitude $\lambda_{consist}$ for both domains.

3.5. Modern Training Techniques

UVCGAN and CycleGAN use outdated GAN training techniques. Hence, we revamp the training procedure with a few modern additions. First, we implement the exponential averaging of the generator weights [20], making the generator less dependent on random fluctuations occurring in the GAN training. Second, we use spectral normalization of the discriminator weights [31], enhancing the stability of the discriminator and training procedure as a whole. Third, we experiment with using unequal learning rates for the generator and discriminator [17], which has been empirically shown to improve model performance. Finally, we replace the generic gradient penalty (GP) with an improved zero-centered GP version [40], developed to promote the GAN's convergence.

4. Experiments

4.1. Datasets

We study the performance of UVCGANv2 on two groups of datasets: the formerly widely used CelebA [27] and Anime [23] datasets and the modern, high-quality CelebA-HQ [20] and AFHQ [8] datasets. More details about these datasets can be found in Appendix A.

CelebA and Anime. The CelebA and Anime datasets have been commonly used to benchmark GAN-based unpaired I2I translation algorithms [23, 32, 42, 49, 52]. We study UVCGANv2 performance on three tasks related to the CelebA and Anime datasets: Male-to-Female translation on the CelebA dataset, Glasses Removal on the CelebA dataset, and Selfie-to-Anime translation on the Anime dataset. Because the CycleGAN setup learns translations in both directions simultaneously, we also get benchmarks in the opposite directions (Female-to-Male, Glasses Addition, and Anime-to-Selfie translations).

CelebA-HQ and AFHQ. To compare the performance of the UVCGANv2 against more recent unpaired I2I translation algorithms, such as EGSDE [48], we also consider the CelebA-HQ and AFHQ datasets. We investigate Male-to-Female translation on CelebA-HQ and three translations, Cat-to-Dog, Wild-to-Dog, and Wild-to-Cat, on AFHQ.

Preprocessing. For a fair comparison with EGSDE [48], we downsize the CelebA-HQ and AFHQ images to 256×256 pixels. To avoid Fréchet inception distance (FID) evaluation inconsistencies associated with variations in the interpolation procedures between different frameworks [35],

we use the Pillow [9] image manipulation library to perform the image resizing with Lanczos interpolation.

4.2. Training

When training our modified UVCGAN implementation, we seek to closely follow the original two-step procedure [42]. The first involves pre-training the generator in a self-supervised way on a task of image inpainting, while the second step is the actual training of the unpaired I2I translation networks, starting from the pre-trained generators.

Generator Pre-training. For each dataset, the generators are pre-trained on image inpainting tasks. This task is set up in a fashion similar to the Bidirectional Encoder Representations from Transformers (BERT) pre-training [11, 42]. For the inpainting task, input images of size 256×256 pixels are tiled into a grid of patches at 32×32 pixels. Then, each patch is masked with a probability of 40%. The masking is performed by zeroing out pixel values. The generator is tasked to recover the original unmasked image from a masked one. More details about this pre-training are available in Appendix B.

Translation Training. The unpaired image translation training is performed for one million iterations using the Adam optimizer. Depending on the dataset, we employ various data augmentations. For the preprocessed datasets, such as CelebA-HQ and AFHQ, only a random horizontal flip is applied. For the Anime and CelebA datasets, we use three augmentations: resize, followed by a random crop of size 256×256 then, finally, a random horizontal flip. Resizing for the Anime dataset is done from 256×256 up to 286×286. For the CelebA dataset, the resizing is done from 178×218 to 256×313.

Hyperparameter Tuning. For each translation, we perform a quick hyperparameter search, exploring the space of the cycle-consistency magnitude λ_{cycle}, magnitude of the zero-centered GP λ_{GP}, magnitude of the consistency loss λ_{consist}, learning rates of the generator and discriminator, and the choice of the batch head (BN versus BSD). We also explore turning the learning rate scheduler on and off. Refer to Appendix B for more details.

5. Results

5.1. Metrics of Realism and Faithfulness

There are two dimensions along which the unpaired I2I style transfer models can be evaluated: *Faithfulness* and *Realism*. Faithfulness captures the degree of similarity between the source and its translated image at an individual level. Realism attempts to estimate the overlap of the distributions of the translated images and the ones in the target.

In terms of realism, image translation quality is commonly judged according to the FID [17] and kernel inception distance (KID) [3] metrics. Both metrics measure the distance between the distributions of the latent Inception-v3 [39] features extracted from samples of the translated and target images. Smaller FID and KID values indicate more realistic images.

Early GAN-based works (e.g., [23, 32, 49, 52]) do not explicitly evaluate the faithfulness of the translation. To the best of our knowledge, there is no widely accepted faithfulness metric available. Some works [48] try to employ simple pixel-wise L_2, peak-signal-to-noise ratio (PSNR), or structural similarity index measure (SSIM) [43] scores to capture the agreement between the source and translation. Yet, it is unclear how well these pixel-wise metrics relate to the perceived image faithfulness, and we explore more advanced alternatives in subsection 5.5.

5.2. Evaluation Protocol

Evaluation protocols differ drastically between different papers (see Appendix C). This makes the direct comparison of the translation quality metrics extremely challenging. For the fairness of comparisons with older works, we follow different evaluation protocols, depending on the dataset.

CelebA and Anime. When evaluating the quality of translation on the CelebA Male-to-Female, CelebA Glasses Removal, and Anime datasets, we use the evaluation protocol of UVCGAN [42], which uniformized FID/KID evaluation across multiple datasets and models, allowing for a simple FID/KID comparison. For the actual FID/KID evaluation, we rely on `torch-fidelity` [33], which provides a validated implementation of these metrics.

The actual evaluation protocol for CelebA and Anime relies only on test splits to perform the FID/KID evaluation. For the CelebA dataset, we use KID subset size of 1000. For the Anime dataset, we use the KID subset size of 50. We use unprocessed images of size 256×256 when evaluating on the Anime dataset. For the CelebA dataset, we apply a simple pre-processing to both domains: resizing the smaller side to 256 pixels then taking a center crop of size 256×256.

CelebA-HQ and AFHQ. EGSDE [48] has evaluated multiple models on the CelebA-HQ and AFHQ datasets under similar conditions. To compare our results to EGSDE, we replicate its evaluation protocol for CelebA-HQ and AFHQ. For the AFHQ dataset, we evaluate FID and KID scores between the translated images of size 256×256 and the target images of size 512×512 from the validation split. For the CelebA-HQ dataset, we evaluate the FID/KID scores between the translated images of size 256×256 and the downsized target images of size 256×256 from the train split. We perform the same channel standardization as EGSDE with $\mu = (0.485, 0.456, 0.406)$ and $\sigma = (0.229, 0.224, 0.225)$. To ensure full consistency, we use the reference evaluation code provided by EGSDE [47]. Appendix E provides results of an alternative evaluation protocol that is uniform across all the datasets.
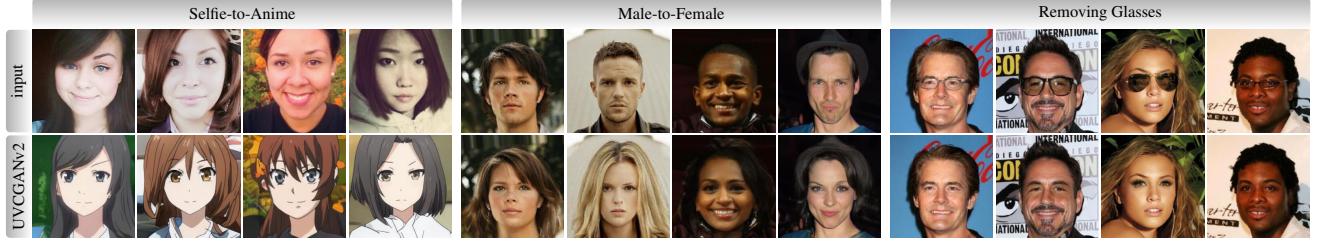
Figure 3. **Sample translations for CelebA and Anime.** Translations produced by UVCGANv2 for three tasks: Selfie-to-Anime, Male-to-Female, and Removing Glasses. The full grid with all benchmarking results can be found in Appendix F.
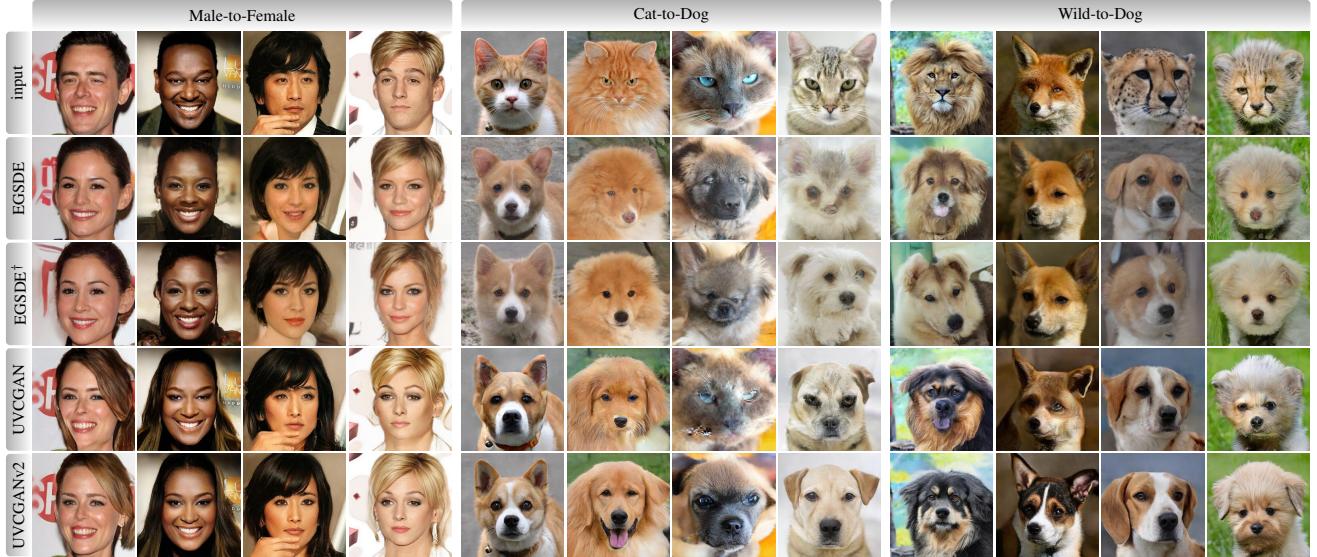

Figure 4. **Sample translations for CelebA-HQ and AFHQ.** Translations for three tasks: Male-to-Female, Cat-to-Dog, and Wild-to-Dog. More translations for these three tasks and those for Wild-to-Cat can be found in Appendix F.

5.3. Quantitative Results

CelebA and Anime. Table 1 shows a comparison of the UVCGANv2 (trained without piwel-wise consistency loss) performance against ACLGAN [49], CouncilGAN [32], CycleGAN [52], U-GAT-IT [23], and UVCGAN [42]. The competitor models' performance is obtained from the UVC-GAN paper [42]. According to Table 1, UVCGANv2 outperforms all competitor models in all translation directions, except Anime-to-Selfie. The degree of improvement ranges from about 5% in terms of FID on the Selfie-to-Anime translation to around 51% on the Male-to-Female translation. Likewise, there is a significant improvement in the KID scores from about 13% on Anime-to-Selfie to 79% on Male-to-Female. This improvement demonstrates the effectiveness of modern additions to the traditional CycleGAN architecture. Figure 3 provides some translation samples, while more can be found in Appendix F.

CelebA-HQ and AFHQ. Table 2 compares the results of the UVCGANv2 evaluation against CUT [34], ILVR [6], SDEdit [30], two versions of the EGSDE [48], and the orig-

inal UVCGAN [42] model. In particular, this table compares two versions of the UVCGANv2: UVCGANv2 and UVCGANv2-C. UVCGANv2-C is a version trained with a pixel-wise consistency loss and $\lambda_{consist} = 0.2$. Here, we train UVCGAN models from scratch (details are provided in Appendix H) and extract the performance metrics of the remaining competitor models from EGSDE [48].

Table 2 shows that UVCGANv2 achieves the best FID scores across all translation tasks. It outperforms EGSDE and UVCGAN by at least 10% on both AFHQ tasks. On the Male-to-Female task, it surpasses EGSDE by a significant amount, 43%. Intriguingly, the original UVCGAN's performance on this task seems on par with UVCGANv2. Yet, UVCGANv2 shows an advantage if we adapt the UVC-GAN's evaluation protocol with 19.3% improvement (FID 24.2 vs 30.0, c.f. subsection E.1).

The addition of the consistency loss allows the UVCGANv2-C model to improve its pixel-wise PSNR and SSIM metrics—but at the expense of the FID score on AFHQ translation. UVCGANv2 and UVCGANv2-C

6

Table 1. **FID and KID scores.** Lower is better.

	Selfie-to-Anime		Anime-to-Selfie	
	FID	KID (×100)	FID	KID (×100)
ACLGAN	99.3	3.22 ± 0.26	128.6	3.49 ± 0.33
CouncilGAN	91.9	2.74 ± 0.26	126.0	2.57 ± 0.32
CycleGAN	92.1	2.72 ± 0.29	127.5	2.52 ± 0.34
U-GAT-IT	95.8	2.74 ± 0.31	**108.8**	1.48 ± 0.34
UVCGAN	79.0	1.35 ± 0.20	122.8	2.33 ± 0.38
UVCGANv2	**75.8**	**1.18 ± 0.28**	113.8	**1.26 ± 0.23**

	Male-to-Female		Female-to-Male	
	FID	KID (×100)	FID	KID (×100)
ACLGAN	9.4	0.58 ± 0.06	19.1	1.38 ± 0.09
CouncilGAN	10.4	0.74 ± 0.08	24.1	1.79 ± 0.10
CycleGAN	15.2	1.29 ± 0.11	22.2	1.74 ± 0.11
U-GAT-IT	24.1	2.20 ± 0.12	15.5	0.94 ± 0.07
UVCGAN	9.6	0.68 ± 0.07	13.9	0.91 ± 0.08
UVCGANv2	**4.7**	**0.14 ± 0.02**	**7.6**	**0.24 ± 0.02**

	Remove Glasses		Add Glasses	
	FID	KID (×100)	FID	KID (×100)
ACLGAN	16.7	0.70 ± 0.06	20.1	1.35 ± 0.14
CouncilGAN	37.2	3.67 ± 0.22	19.5	1.33 ± 0.13
CycleGAN	24.2	1.87 ± 0.17	19.8	1.36 ± 0.12
U-GAT-IT	23.3	1.69 ± 0.14	19.0	1.08 ± 0.10
UVCGAN	14.4	0.68 ± 0.10	13.6	0.60 ± 0.08
UVCGANv2	**10.6**	**0.27 ± 0.06**	**11.3**	**0.34 ± 0.07**

Table 2. **FID, PSNR, SSIM, and LPIPS scores.**

	Male-to-Female (CelebA-HQ)			
	FID↓	PSNR↑	SSIM↑	LPIPS↑
CUT	46.61	19.87	0.74	-
ILVR	46.12	18.59	0.510	-
SDEdit	49.43	20.03	0.572	-
EGSDE	41.93	20.35	0.574	0.162
EGSDE[†]	30.61	18.32	0.510	0.159
UVCGAN	17.63	19.30	**0.753**	**0.194**
UVCGANv2	17.65	19.44	0.681	0.188
UVCGANv2-C	**17.34**	**21.18**	0.738	0.190

	Cat-to-Dog			
	FID↓	PSNR↑	SSIM↑	LPIPS↑
CUT	76.21	17.48	0.601	-
ILVR	74.37	17.77	0.363	-
SDEdit	74.17	19.19	0.423	-
EGSDE	65.82	**19.31**	0.415	0.199
EGSDE[†]	51.04	17.17	0.361	0.196
UVCGAN	69.33	18.36	**0.683**	**0.227**
UVCGANv2	**44.76**	15.55	0.562	0.221
UVCGANv2-C	52.48	18.30	0.638	0.221

	Wild-to-Dog			
	FID↓	PSNR↑	SSIM↑	LPIPS↑
CUT	92.94	17.2	0.592	-
ILVR	75.33	16.85	0.287	-
SDEdit	68.51	17.98	0.343	-
EGSDE	59.75	18.14	0.343	0.193
EGSDE[†]	50.43	16.40	0.300	0.190
UVCGAN	78.44	17.67	**0.675**	**0.224**
UVCGANv2	**45.56**	15.59	0.551	0.217
UVCGANv2-C	55.61	**18.65**	0.631	0.219

achieve competitive SSIM scores but lose in terms of the PSNR ratio to the other models. However, as previously noted [46], pixel-wise measures PSNR and SSIM are not good metrics to judge perceptual image faithfulness. Overall, the gains in SSIM and PNSR scores provided by the consistency loss to UVCGANv2-C do not seem to outweigh the associated FID losses.

Finally, it may be instructive to examine the diversity of the translated images. We compare the diversity of the generated images according to a pairwise LPIPS distance [46], following an approach similar to [26]. To calculate the LPIPS distance, we use a VGG-based implementation (v0.0) in a consecutive pair mode. Table 2 shows that UVCGANv2 produces a much greater diversity of generated images compared to the EGSDE variants, yet it is slightly less diverse compared to UVCGAN. It also indicates that the presence of the pixel-wise consistency loss leads to a small but repeatable increase in image diversity.

5.4. Ablation Study

Table 3 summarizes UVCGANv2 ablation results on the Male-to-Female translation of CelebA. To produce this table, we start with the final UVCGANv2 configuration and make one of following modifications separately: (a) disable style modulation in the generator; (b) disable batch head of the discriminator; (c) revert to UVCGAN training setup:

Table 3. **Ablation Study of UVCGANv2 on CelebA.**

	Male-to-Female		Female-to-Male	
	FID	KID (×100)	FID	KID (×100)
UVCGAN	9.6	0.68 ± 0.07	13.9	0.91 ± 0.08
UVCGANv2	**4.7**	**0.14 ± 0.02**	**7.6**	**0.24 ± 0.02**
(a) w/o New Gen.	8.1	0.53 ± 0.07	11.1	0.64 ± 0.07
(b) w/o New Disc.	5.5	0.21 ± 0.03	8.6	0.31 ± 0.03
(c) w/o New Tech.	5.6	0.25 ± 0.04	10.0	0.51 ± 0.05

linear scheduler, full GP, and disable generator averaging.

According to Table 3, the generator modifications (a) account for the majority of the performance improvement. Removing these modifications degrades the FID score of the Male-to-Female translation from 4.7 to 8.1. The discriminator modifications (b) provide a significant but relatively smaller improvement in the I2I performance. The new training setup (c) produces an improvement somewhere in between the generator and discriminator modifications.

5.5. Toward Better Faithfulness Measures

Pixel-wise image similarity measures (such as L_2, PSNR, and SSIM) have been shown [46] to be weakly cor-

Table 4. **Faithfulness scores.**

| | Male-to-Female (CelebA-HQ) | | | |
	CLIP↑	LPIPS↓	I-L_2↓	L_2↓
EGSDE	0.637	0.116	13.93	**43.41**
EGSDE†	0.600	0.131	15.22	54.50
UVCGAN	**0.734**	**0.089**	**13.12**	54.08
UVCGANv2	<u>0.723</u>	0.101	<u>13.36</u>	64.23
UVCGANv2-C	0.720	<u>0.093</u>	13.47	<u>47.94</u>

| | Cat-to-Dog | | | |
	CLIP↑	LPIPS↓	I-L_2↓	L_2↓
EGSDE	0.762	0.165	16.65	**49.20**
EGSDE†	0.749	0.181	16.74	62.44
UVCGAN	**0.822**	**0.126**	16.44	<u>54.45</u>
UVCGANv2	0.802	0.150	**16.31**	77.94
UVCGANv2-C	<u>0.809</u>	<u>0.140</u>	<u>16.43</u>	56.83

| | Wild-to-Dog | | | |
	CLIP↑	LPIPS↓	I-L_2↓	L_2↓
EGSDE	0.672	0.176	15.18	**56.65**
EGSDE†	0.649	0.192	15.41	68.46
UVCGAN	**0.765**	**0.124**	15.14	61.05
UVCGANv2	0.720	0.150	**14.75**	81.75
UVCGANv2-C	<u>0.740</u>	<u>0.136</u>	<u>14.81</u>	<u>58.77</u>

related with the human perception of similarity. However, they currently are being used [48] as a faithfulness metric in the area of unpaired I2I translation. In this section, we explore using CLIP [36] and LPIPS [46] scores as similarity measures because they have been shown to match human perception. In addition, it may be more natural to build a similarity measure reusing Inception-v3 features that are already employed in the FID calculation. This will avoid introducing new dependencies, parameter choices, and other sources of inconsistency. Thus, as another faithfulness metric, we propose an *Inception-v3 L_2 distance* (I-L_2) formed between the Inception-v3 features.

Table 4 compares the faithfulness scores according to CLIP, LPIPS, I-L_2, and pixel-wise L_2. It indicates that UVCGANv2 variants outperform EGSDE variants on the perceptual metrics. On the other hand, EGSDE shows superiority in the pixel-wise scores. At the same time, UVC-GAN demonstrates better perceptual faithfulness compared to UVCGANv2. Given that UVCGANv2 has better realism than UVCGAN, this likely demonstrates an example of a realism-faithfulness trade-off. The trends of the proposed I-L_2 roughly match those of the more complex CLIP and LPIPS measures, illustrating the feasibility of an I-L_2 as a simpler similarity measure.

Fundamentally, one still may wonder if any of these metrics are "proper" measurements of faithfulness? When measuring faithfulness, one should differentiate between the *domain-contrastive* features (changing between domains)

and *domain-consistent* features (expected to be preserved). For example, in the CelebA-HQ dataset, males tend to have shorter hair compared to females. Thus, we expect a Male-to-Female I2I algorithm will increase hair length on average (domain-contrastive feature). On the other hand, image background, skin color, facial expression, and facial orientation are approximately the same in both domains. Therefore, an I2I algorithm is expected to preserve these features (domain-consistent features). We argue a proper faithfulness metric should pay the most attention to domain-consistent features and be indifferent to the domain-contrastive features.
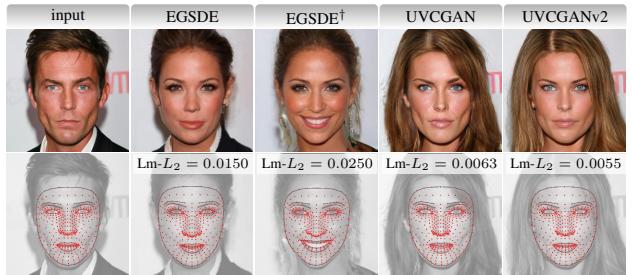

Figure 5. Examples of face landmark distance (Lm-L_2) between an input and a translated image. (Appendix D)

As an initial attempt, we propose an approximation to such a metric: similarity of face landmarks [5, 10, 29]. Face landmarks are key points on a human face, capturing information about its position, orientation, structure, expression, etc. That is, capturing those features expected to be largely preserved during the Male-to-Female translation. Therefore, one may consider forming an Euclidean distance (Lm-L_2) between the landmark locations of the source and translated images (Figure 5). Better faithfulness metrics also will consider other domain-consistent features. This development is left for future work.

6. Conclusion

This work revisits the UVCGAN model, built upon the classic CycleGAN architecture. It demonstrates that source-driven style modulation and batch-statistics-aware discriminator are effective techniques to improve the model performance. Our UVCGANv2 has been extensively benchmarked on four datasets with nine translation directions. Results show our model can achieve superior performance in terms of FID scores. At the same time, we note the absence of a proper faithfulness metric in the area of unpaired I2I translation and general inconsistency of the evaluation procedures. As such, developing such a measure and uniformizing evaluation will be highly beneficial for future unpaired I2I methods development.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 3

[2] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34:22614–22627, 2021. 1

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5, 16

[4] M Boulanger, Jean-Claude Nunes, H Chourak, A Largent, S Tahri, O Acosta, R De Crevoisier, C Lafond, and A Barateau. Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review. *Physica Medica*, 89:265–281, 2021. 1

[5] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollar. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013. 8

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 1, 2, 6

[7] 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. 1

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 1, 4, 11, 13

[9] Alex Clark. Pillow, python imaging library (fork). 5, 16

[10] Clement Creusot, Nick Pears, and Jim Austin. 3d face landmark labelling. In *Proceedings of the ACM workshop on 3D object retrieval*, pages 27–32, 2010. 8

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5, 11

[12] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 1

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[15] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[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. *Advances in neural information processing systems*, 30, 2017. 4, 5

[18] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 1

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3, 4, 11

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 3

[22] Junho Kim. GAN_Metrics-Tensorflow, simple tensorflow implementation of metrics for gan evaluation. https://github.com/taki0112/GAN_Metrics-Tensorflow. 12

[23] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 1, 2, 4, 5, 6, 11, 12

[24] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 1

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[26] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16377–16386, 2021. 7

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 4, 11

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei

Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. 8, 13

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2, 6

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 4

[32] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7860–7869, 2020. 2, 4, 5, 6, 11

[33] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738. 5, 16

[34] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 1, 2, 6, 13

[35] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 4, 13, 16

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[37] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 1

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[40] Hoang Thanh-Tung, Truyen Tran, and Svetha Venkatesh. Improving generalization and stability of generative adversarial networks. *arXiv preprint arXiv:1902.03984*, 2019. 1, 4

[41] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. uvcgan, 2022. 17

[42] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 702–712, 2023. 1, 2, 3, 4, 5, 6, 11, 13, 17

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 4

[45] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7, 8, 13

[47] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. EGSDE energy-guided stochastic differential equations. https://github.com/ML-GSAI/EGSDE. 5

[48] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022. 1, 2, 4, 5, 6, 8, 13, 16

[49] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 800–815. Springer, 2020. 2, 4, 5, 6

[50] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16407–16417, 2021. 2

[51] Wanfeng Zheng, Qiang Li, Guoxin Zhang, Pengfei Wan, and Zhongyuan Wang. Ittr: Unpaired image-to-image translation with transformers. *arXiv preprint arXiv:2203.16015*, 2022. 2

[52] 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. 1, 2, 4, 5, 6, 12

[53] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1

A. Datasets

In this section, we provide additional details about the datasets used in the main paper.

CelebA [27]. The datasets for the Glasses Removal and Male-to-Female tasks are derived from the original CelebA dataset. For a fair comparison with older models [42], we used the pre-processed versions of Glasses Removal and Male-to-Female datasets provided by CouncilGAN [32]. The CelebA dataset is made of images of size 178×218 pixels. The train split of the Glasses Removal task contains about 11K images with glasses and 152K without. The Male-to-Female dataset has about 68K males and 95K females. The test parts of the Glasses Removal and Male-to-Female datasets contain about 3K images with glasses and 37K images without glasses, and 16K males and 24K females respectively.

Anime-to-Selfie [23]. The training split of the Anime-to-Selfie dataset contains 3400 Selfie images and 3400 Anime images. The test part contains just 100 samples from each domain. All images of this dataset have a size of 256×256 pixels.

CelebA-HQ [20]. The CelebA-HQ dataset has around 10K images of males and about 18K images of females in the train split. The test split contains 1000 male and 1000 female images. The size of a CelebA-HQ image is 1024×1024 pixels.

AFHQ [8]. The AFHQ dataset has around 5.2K cat, 4.7K dog, and 4.7K wildlife images in the train split, and 500 images of each in the test split. The images of the AFHQ dataset have a size of 512×512 pixels. There are two versions of the AFHQ dataset provided by StarGANv2 [8]. We use version 1 to be consistent with previous models.

B. Training Details

In this section, we expand on the generator pre-training, I2I translation training, HP optimization setup, the final model configurations, and the ablation study of UVC-GANv2.

Generator Pre-Training. The pre-training of the generators was performed in a BERT-like fashion [11] on an image inpainting pretext task.

To construct the image inpainting task, input images of size 256×256 pixels are tiled into a grid of patches of 32×32 pixels. Then, each patch is randomly masked with a probability of 40%. The masking is performed by zeroing out pixel values. The generator is tasked to recover the original unmasked image from a masked one.

For the pre-training, we use the AdamW optimizer together with a cosine learning rate annealing (with restarts). We set the initial learning rate to $5 \times 10^{-3} \times$ (batch size$/512$), and the weight decay factor to 0.05. The scheduler completes 5 annealing cycles during the pre-training.

We apply several data augmentations, such as random rotation (± 10 degree), random horizontal flip ($p = 0.5$), and color jitter (± 0.2 shift in brightness, contrast, saturation, and hue).

We pre-train the generators for 500 epochs for CelebA-HQ and AFHQ. Due to the small size of the Anime dataset, we run the pre-training for 2500 epochs. On the contrary, due to the large size of the CelebA dataset, we run the pre-training for 500 epochs, but limit the number of samples per epoch to $32,768$. All the pre-trainings are performed with a batch size of 64.

Image Translation Training. We train the unpaired I2I translation models by closely following the procedure of UVCGAN [42]. We use the Adam optimizer without weight decay. The training is performed for 1 million iterations. We experiment with using either a constant learning rate or applying a linear scheduler. If the linear scheduler is used, then the learning rate is maintained constant for the first 500K iterations, and then linearly annealed to zero during the subsequent 500K iterations.

We keep the batch size equal to 1 during the training. For consistency with ProGAN [20] we keep the sizes of the image caches at 3. This size effectively provides the batch head with 4 samples to estimate the batch statistics. To stabilize the generator, we apply an exponential weight averaging to the generator with a momentum of 0.9999.

Hyperparameter Exploration. When performing the final training, we explored the following grid of hyperparameters:

- Magnitude of the cycle-consistency loss λ_{cyc}: $\{5, 10\}$.

- Magnitude of the zero-centered gradient penalty λ_{GP}: $\{0.001, 0.01, 1\}$.

- Batch Head type: Batch Normalization (BN) vs Batch Standard Deviation (BSD).

- Generator's and Discriminator's learning rates:

  1. Equal learning rates of 1×10^{-4}.

  2. Unequal learning rates, with the learning rate of the discriminator of 1×10^{-4} and the learning rate of the generator of 5×10^{-5}.

The hyperparameter explorations were performed while keeping the magnitude of the consistency loss $\lambda_{\mathrm{consist}}$ equal to zero. The grid of hyperparameters above was suggested by the previous rough HP exploration.

For the AFHQ Cat-to-Dog, Wild-to-Dog, and CelebA-HQ Male-to-Female datasets, we have run a second hyperparameter exploration, studying the effect of the magnitude of the consistency loss $\lambda_{\mathrm{consist}}$ on the I2I performance. We have tried the following values of $\lambda_{\mathrm{consist}}$: $\{0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$.
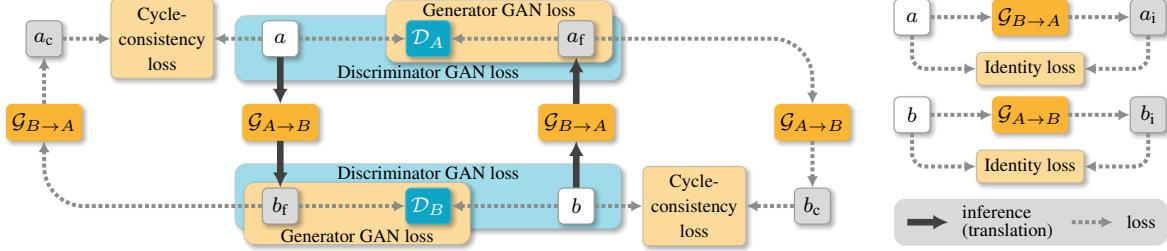
Figure 6. **CycleGAN framework.** The CycleGAN [52] consists of two pairs of GANs, $(\mathcal{G}_{A \to B}, \mathcal{D}_B)$ and $(\mathcal{G}_{B \to A}, \mathcal{D}_A)$. The discriminators try to distinguish translations from real images, while the generators (or translators) seek to produce realistic translations that are also consistent with the input. The consistency is enforced by the cycle-consistency loss and (optional) identity loss. Here, we use a to denote an image from domain A, b as an image from domain B, $(*)_f$ is a fake image (a translation), $(*)_c$ notes a cyclic reconstruction, and $(*)_i$ represents an identity reconstruction (when identity losses are used, $\mathcal{G}_{A \to B}|_B$ and $\mathcal{G}_{B \to A}|_A$ are encouraged to be identity maps).

B.1. Final Configurations

Table 5. **Best Training Configurations.**

Dataset	lr_{gen}	λ_{GP}	λ_{cyc}	B. Head
Anime to Selfie	5×10^{-5}	0.01	10	BN
Male to Female	1×10^{-4}	0.01	10	BSD
Glasses Removal	5×10^{-5}	0.01	10	BSD
HQ Male to Female	1×10^{-4}	1.0	5	BSD
Cat to Dog	1×10^{-4}	1.0	5	BN
Wild to Dog	1×10^{-4}	1.0	5	BN
Cat to Wild	5×10^{-5}	1.0	5	BN

For all the translation tasks, UVCGANv2 achieves the best performance when the learning rate scheduler is not used. Table 5 summarizes the final hyperparameter configurations (generator's learning rate, magnitudes of the gradient penalty and the cycle consistency loss, and the choice of batch head) that provide the best performance per translation task.

Generally, the high-quality datasets (CelebA-HQ and AFHQ) favor stronger values of the gradient penalty term $\lambda_{GP} = 1$, compared to the lower-resolution datasets (Anime-to-Selfie and CelebA), favoring $\lambda_{GP} = 0.01$. Other patterns of hyperparameters can be observed in the table. However, their impact on the model's performance is relatively small compared to λ_{GP}.

We should note, that due to the instability associated with GANs training, some of the best values in Table 5 may be due to random fluctuations.

B.2. Ablation Study

In this section we provide a more detailed ablation study of UVCGANv2. Table 6 summarizes UVCGANv2 ablation results on the Male-to-Female translation of CelebA. To produce this table we start with the final UVCGANv2 configuration and make one of following modifications separately: (a) disable style modulation in the generator; (b)

Table 6. **Ablation Study of UVCGANv2 on CelebA.** "+" indicates that an option is added to the final UVCGANv2 configuration. "−" indicates that an option is removed from UVCGANv2.

	Male-to-Female		Female-to-Male	
	FID	KID ($\times 100$)	FID	KID ($\times 100$)
UVCGAN	9.6	0.68 ± 0.07	13.9	0.91 ± 0.08
UVCGANv2	**4.7**	$\mathbf{0.14 \pm 0.02}$	**7.6**	$\mathbf{0.24 \pm 0.02}$
(a) − New Gen.	8.1	0.53 ± 0.07	11.1	0.64 ± 0.07
(b) − New Disc.	5.5	0.21 ± 0.03	8.6	0.31 ± 0.03
(c) − New Tech.	5.6	0.25 ± 0.04	10.0	0.51 ± 0.05
(d) − SN	4.7	0.14 ± 0.02	7.7	0.25 ± 0.03
(e) + Sched	6.3	0.35 ± 0.05	9.5	0.47 ± 0.05
(f) − Avg.	9.8	0.71 ± 0.05	14.2	0.91 ± 0.07

disable batch head of the discriminator; (c) use UVCGAN training setup: linear scheduler, full GP, and disable generator averaging; (d) disable spectral normalization (SN); (e) add linear schedule; (f) remove exponential averaging of the generator weights.

Table 6 shows, that the removal of the spectral normalization (d) decreases the model performance, but the effect is negligible. This indicates that the Spectral Normalization might be redundant and can be removed.

Each of the items (e) and (f) of Table 6 may suggest that either the scheduler or exponential averaging of the generator weights is detrimental to the model's performance. However, this is not the case, since the effects of (d) and (e) are entangled and mutually balancing. Individual modifications to (d) or (e) destroy the balance and produce large changes in the model's performance. These changes are not indicative of the effects of the joint modifications as shown by (c).

C. Remarks on Metric Evaluation Consistency

Inconsistency of unpaired I2I evaluation procedures is a widespread problem. For example, some works (e.g. [23]) roll out their own FID evaluation code [22] and report

the so-called "mean" FID and KID scores, where "mean" means a weighted average of the actual FID/KID scores and various other metrics. Some other works [35] choose different image resizing algorithms, creating a noticeable discrepancy in the reported FID scores.

Another source of FID/KID score inconsistency is the difference in the evaluation protocols. For instance, works like [42] prefer to evaluate FID scores only on images of the test split, yet others [8] evaluate FID scores between translated test images and target images of the train split. Likewise, there is a difference in whether any pre-processing is used in the FID/KID evaluation. For example, one can evaluate FID scores between translated images with the pre-processing and target images without pre-processing [34], or one can apply the pre-processing step to both translated and target images [48]. Moreover, the pre-processing procedures may differ between different works.

To uniformize FID/KID evaluation procedures, we propose a consistent evaluation protocol in Appendix E.

D. Metrics of Faithfulness to the Source

In this section, we provide more discussion on the two faithfulness metrics we proposed: Inception v3 L_2 (I-L_2) and Landmark L_2 (Lm-L_2).

Inception v3 L_2 Distance: Here we present more examples to illustrate that I-L_2 (defined as the L_2 distance between the latent Inception-v3 features) may be a more appropriate measurement of faithfulness than pixel-level L_2. In Table 7, we select translations according to two types of criteria. Denote a translation produced by EGSDE as E and that by UVCGANv2 as U.

Type 1: I-$L_2(E) \approx$ I-$L_2(U)$ and $L_2(U) - L_2(E) > 15$.
Type 2: I-$L_2(E) -$ I-$L_2(U) > 3$.

Type 1 is designed to show what contributes to lower pixel-wise L_2 while I-L_2 scores are similar. **Type 2** selects examples with large I-L_2 difference and helps readers to judge if I-L_2 correlates with their own judgment on the similarity to the source (i.e. which pairs look more like siblings.)

We list eight categories of perceived image faithfulness (PIF) in the legend such as background, bone structure, facial expression, and so on. For each translated image, we indicate which categories it outperforms that generated by the other model. For example, for the input in Type 1 row 1 left, EGSDE preserves the hairstyle and color (PIF=7) better than UVCGANv2, but the UVCGANv2 translation maintains a sharper background (1), preserves the bone structure (2) and expression (3) better, and exhibits more similarity in apparent age (4).

Type-1 examples suggest that pixel-level L_2 is an inappropriate measurement for semantic consistency as UVC-GANv2 translations manage to capture characterizing fea-

tures (such as a bone structure) even with high pixel-level L_2. In fact, the high pixel-level L_2 of UVCGANv2 translations is often a result of benign modifications such as the elongation of dark hair on a light background (e.g. Type 1 row 5 right) or a slight overall shift to a warmer hue (e.g. Type 1 row 2 right).

On the contrary, the Type-2 examples suggest that I-L_2, the L_2 on Inception latent features, might be a better measurement of semantic consistency. While EGSDE fails to maintain features such as background, facial expression (neural v.s. smile), eye movement, and prominent bone structure and produces over-generalized translations, UVC-GANv2 translations with significantly lower I-L_2 manage to preserve those features and appear more individualized.

These examples illustrate that the pixel-wise L_2 faithfulness metric may be in poor agreement with a human judgment of image faithfulness. They also point to a possibility that the I-L_2 distances, based on deep features of Inception-v3, may better capture the perceived image faithfulness. Such observations mirror the conclusion of [46], about the effectiveness of deep features as perceptual metrics.

However, we stress again, that the main purpose of this paper is to improve the performance of the classic Cycle-GAN architecture, not the development of better faithfulness metrics. While this section points to a possibility of I-L_2 being a better faithfulness metric, a full-scale investigation needs to be conducted to conclusively establish this. We leave such a study for future work.

Landmark L_2 Distance: We start with defining Lm-L_2. Let $\mathcal{L}^i = \left(l_1^i, \ldots, l_N^i \right)$ be the landmarks of an input face and let $\mathcal{L}^t = (l_1^t, \ldots, l_N^t)$ be those of a translated face. Lm-L_2 is defined as follows:

$$\text{Lm-}L_2 \left(\mathcal{L}^i, \mathcal{L}^t \right) = \frac{1}{N} \sum_{i=1}^{N} \left\| l_i^i - l_i^t \right\|_2.$$

For this study, we calculate face landmarks with mediapipe (v0.10.1) [29] with $N = 468$ and $l_i = (x, y, z)$ being the location of the landmark relative to the image. From Table 8, we can see that UVCGANv2 is more faithful with respect to Lm-L_2. Find more examples of face landmarks with their Lm-L_2 in Table 9.

E. Toward Consistent FID Evaluation

The evaluation protocols used in the paper for CelebA-HQ and AFHQ are provided by EGSDE [48]. Being ad-hoc, these protocols lack consistency and differ significantly depending on the dataset. A variety of different evaluation protocols makes the evaluation of the unpaired I2I methods rather quirky and error-prone.

As a step toward consistent FID evaluation, we extend a more uniform evaluation protocol of UVCGAN [42] toward

Table 7. **Comparing EGSDE and UVCGANv2 translations with L_2 and I-L_2 scores.**

	EGSDE	input	UVCGANv2	EGSDE	input	UVCGANv2
Type 1	$L_2 = 35.7$, I-$L_2 = 12.6$, PIF $= 7$		$L_2 = 70.5$, I-$L_2 = 10.9$, PIF $= 1, 2, 3, 4$	$L_2 = 40.9$, I-$L_2 = 11.4$, PIF $= 7$		$L_2 = 59.0$, I-$L_2 = 11.5$, PIF $= 2, 3, 6$
	$L_2 = 37.6$, I-$L_2 = 13.2$, PIF $=$		$L_2 = 54.3$, I-$L_2 = 11.1$, PIF $= 2, 3, 4$	$L_2 = 48.9$, I-$L_2 = 14.8$, PIF $=$		$L_2 = 75.2$, I-$L_2 = 11.4$, PIF $= 1, 3, 5, 6$
	$L_2 = 39.9$, I-$L_2 = 10.6$, PIF $= 7$		$L_2 = 64.7$, I-$L_2 = 10.8$, PIF $= 2, 3$	$L_2 = 43.6$, I-$L_2 = 12.6$, PIF $= 7, 8$		$L_2 = 62.0$, I-$L_2 = 11.9$, PIF $= 2, 3, 4$
	$L_2 = 41.7$, I-$L_2 = 13.3$, PIF $=$		$L_2 = 71.2$, I-$L_2 = 10.5$, PIF $= 2, 3, 4$	$L_2 = 40.9$, I-$L_2 = 12.8$, PIF $= 8$		$L_2 = 69.0$, I-$L_2 = 11.9$, PIF $= 1, 2, 4, 6$
	$L_2 = 47.8$, I-$L_2 = 10.2$, PIF $= 7, 8$		$L_2 = 95.9$, I-$L_2 = 11.7$, PIF $= 3, 4, 5, 6$	$L_2 = 43.4$, I-$L_2 = 12.8$, PIF $=$		$L_2 = 80.9$, I-$L_2 = 10.5$, PIF $= 1, 2, 3$
	$L_2 = 41.0$, I-$L_2 = 11.9$, PIF $=$		$L_2 = 59.6$, I-$L_2 = 11.5$, PIF $= 1, 2, 4, 6$	$L_2 = 43.9$, I-$L_2 = 10.6$, PIF $=$		$L_2 = 71.9$, I-$L_2 = 10.6$, PIF $= 2, 3, 4, 8$
	$L_2 = 46.2$, I-$L_2 = 11.2$, PIF $= 8$		$L_2 = 81.5$, I-$L_2 = 10.3$, PIF $= 1, 2, 4, 6$	$L_2 = 46.2$, I-$L_2 = 11.0$, PIF $=$		$L_2 = 66.4$, I-$L_2 = 11.1$, PIF $= 2, 3, 4, 8$
Type 2	$L_2 = 43.6$, I-$L_2 = 13.3$, PIF $= 7$		$L_2 = 69.3$, I-$L_2 = 10.3$, PIF $= 2, 3, 5$	$L_2 = 45.6$, I-$L_2 = 18.0$, PIF $= 7$		$L_2 = 113.7$, I-$L_2 = 14.5$, PIF $= 2, 3, 5, 6$
	$L_2 = 40.5$, I-$L_2 = 19.2$, PIF $=$		$L_2 = 48.1$, I-$L_2 = 15.5$, PIF $= 2, 3$	$L_2 = 49.7$, I-$L_2 = 16.4$, PIF $= 6$		$L_2 = 38.2$, I-$L_2 = 11.6$, PIF $= 1, 2, 3, 5$
	$L_2 = 45.1$, I-$L_2 = 17.0$, PIF $= 7$		$L_2 = 43.2$, I-$L_2 = 13.7$, PIF $= 1, 2, 3, 6$	$L_2 = 40.0$, I-$L_2 = 14.9$, PIF $= 7$		$L_2 = 50.0$, I-$L_2 = 9.1$, PIF $= 1, 2, 3, 6$

Categories of perceived image faithfulness (PIF):
1. background   2. bone structure   3. expression   4. apparent age   5. eye color   6. eyebrows   7. hair color   8. hair texture

Table 8. **Faithfulness by Landmark L_2 (Lm-L_2).**

	EGSDE	EGSDE†	UVCGAN	UVCGANv2	UVCGANv2-C
Lm-$L_2\downarrow$	0.0188	0.0227	0.0129	0.0152	0.0163

the AFHQ and CelebA-HQ datasets. The UVCGANv2 evaluation results under this consistent evaluation protocol are shown in Table 10. The consistent evaluation protocol uses only test splits (or validation splits if the test ones are not available) of each dataset to assess the quality of image

Table 9. **Face landmarks with Lm-L_2**

Table 10. **Consistent FID and KID scores.**

| | Female to Male | | Male to Female | |
	FID↓	KID ($\times 100$)	FID↓	KID ($\times 100$)
UVCGANv2	29.7	0.41 ± 0.18	24.2	0.20 ± 0.15
	Dog to Cat		Cat to Dog	
	FID	KID ($\times 100$)	FID	KID ($\times 100$)
UVCGANv2	24.8	0.23 ± 0.13	44.2	0.76 ± 0.23
	Dog to Wild		Wild to Dog	
	FID	KID ($\times 100$)	FID	KID ($\times 100$)
UVCGANv2	18.7	0.15 ± 0.14	44.7	0.68 ± 0.23
	Cat to Wild		Wild to Cat	
	FID	KID ($\times 100$)	FID	KID ($\times 100$)
UVCGANv2	12.1	0.01 ± 0.09	21.2	0.20 ± 0.13

translation.

The evaluation protocol begins with pre-processing all the datasets in a consistent manner. The pre-processing step resizes images from their original size down to 256×256 pixels (the same image size as is used for model training and inference). To avoid FID score inconsistencies created by aliasing artifacts [35] we rely on the `Pillow` library [9] and Lanczos interpolation method.

Once the data pre-processing and image translation are done, the actual evaluation can begin. To perform the FID/KID score computation we use a `torch-fidelity` package [33], which provides a validated implementation of these metrics. The KID evaluation procedure depends on a free parameter – the KID subset size. In this section, we choose the KID subset size of 100 for all the datasets.

The suggested evaluation protocol differs in a number of ways from the evaluation protocols of the AFHQ and CelebA-HQ datasets of EGSDE. It differs from the ad-hoc CelebA-HQ evaluation protocol [48] because the latter compares FID scores between samples of validation and train splits, while the consistent version only uses validation split. The consistent evaluation protocol is also different from the ad-hoc version of the AFHQ one, which performs FID evaluation between translated images of size 256×256 and target images of size 512×512. The consistent protocol always uses pre-processed images of size 256×256.

E.1. CelebA-HQ Consistent Evaluation

In this section, we elaborate on the CelebA-HQ FID scores obtained with the consistent protocol. Table 10 shows the FID evaluation scores of UVCGANv2 according to the consistent evaluation protocol. A few observations one can draw comparing the results from this table to the ad-hoc evaluations, presented in Table 2 of the main paper. In particular, there is a good agreement of the FID scores on the AFHQ dataset, since the ad-hoc evaluation procedure is

Table 11. **Model Performance versus magnitude of the pixel-wise consistency loss.**

| λ_{consist} | Male-to-Female | | Cat-to-Dog | |
	FID↓	$L_2 \downarrow$	FID↓	$L_2 \downarrow$
0	24.2	62.6	44.2	77.9
0.01	24.9	61.2	44.5	76.7
0.1	24.8	50.6	45.7	64.9
0.2	25.1	47.9	51.8	56.8
0.4	27.3	43.3	59.1	50.6
0.6	29.7	41.0	71.3	47.5
0.8	32.0	39.1	77.0	46.1
1.0	33.1	37.7	81.2	44.9

Table 12. **Consistent UVCGAN evaluation on CelebA-HQ.**

| | Female to Male | | Male to Female | |
	FID↓	KID ($\times 100$)	FID↓	KID ($\times 100$)
EGSDE	-	-	68.3	6.10 ± 0.48
EGSDE†	-	-	53.9	4.31 ± 0.35
UVCGAN	34.6	0.78 ± 0.40	30.0	0.74 ± 0.32
UVCGANv2	**29.7**	$\mathbf{0.41 \pm 0.18}$	**24.2**	$\mathbf{0.20 \pm 0.15}$

very similar to the consistent one. However, there is a large discrepancy in the FID scores on CelebA-HQ.

Several factors contribute to the discrepancy of the FID scores on the CelebA-HQ dataset. First, the ad-hoc method evaluates the FID scores against the train part of the dataset, while the consistent method uses validation parts. This explains the large difference in the scale of the FID values [3]. Second, the consistent evaluation uses a Lanczos interpolation method, less affected by aliasing artifacts. Finally, the consistent version relies on a validated FID calculation code of `torch-fidelity`, while the ad-hoc method of EGSDE uses a custom FID implementation.

Due to the large differences in the evaluation protocols on the CelebA-HQ dataset, it may be instructive to compare FID scores on CelebA-HQ, obtained with the consistent protocol. Table 12 compares the performance of EGSDE, UVCGAN, and UVCGANv2 under the consistent evaluation mode. To a large extent, the behavior of the FID scores in Table 12 matches the behavior of the FID scores of the ad-hoc protocol. At the same time, according to the consistent evaluation protocol, UVCGANv2 outperforms UVC-GAN by $15 - 20\%$ in the FID scores. This is in contrast to the ad-hoc evaluation procedure of Table 2 (the main part of the paper) which does not reflect any difference in the performance.

F. Additional Translation Samples

In this section, we provide additional translation samples to facilitate visual comparison of image quality. Table 13 and 14 demonstrate samples on the Anime dataset. Table 15

and 16 provide gender swap samples on the CelebA dataset. Table 17 and 18 show eyeglasses removal and addition samples on the CelebA dataset. Finally, Table 19, 20, and 21 provide samples on the AFHQ dataset, and Table 22, Male-to-Female samples on the CelebA-HQ dataset.

G. Effects of the Pixel-Wise Consistency Loss

The main part of the paper compares models trained with two settings of the pixel-wise consistency loss. The UVCGANv2 model having $\lambda_{consist} = 0$ and UVCGANv2-C model with $\lambda_{consist} = 0.2$. In this section, we show an ablation of the $\lambda_{consist}$ values and their effect on the model performance.

Table 11 demonstrates the effect of different values of $\lambda_{consist}$ on the UVCGANv2 model realism (as measured by the FID scores) and pixel-wise faithfulness (as measured by the pixel-wise L_2 distance).

As one might expect, the increase in $\lambda_{consist}$ is accompanied by an improvement in pixel-wise image faithfulness and a decrease in image realism. Values of $\lambda_{consist}$ below 0.2 allow one to significantly improve pixel-wise image faithfulness at the expense of a modest loss of image realism. Further increases in $\lambda_{consist}$ produce larger improvements in pixel-wise faithfulness, but also lead to significant decreases in image realism.

Additionally, Table 11 demonstrates that the trade-offs of image realism to pixel-wise faithfulness are not uniform across the datasets. High values of $\lambda_{consist}$ allow one to achieve rather large improvements in pixel-wise faithfulness on the Male-to-Female task at a relatively small loss of image realism. On the other hand, Cat-to-Dog translation is subject to a catastrophic loss of image realism with the increase of $\lambda_{consist}$.

H. UVCGAN Training Procedure

In this appendix, we provide details of the training setup of UVCGAN models for the high-quality datasets (CelebA-HQ and AFHQ). The UVCGAN paper [42] did not study these datasets, therefore we trained the corresponding models (CelebA-HQ Male-to-Female, Cat-to-Dog, and Wildlife-to-Dog) from scratch.

Before training UVCGAN I2I translation generators, we performed their self-supervised pre-training. While performing the pre-training, we closely followed the original procedure [42]. For reference, this pre-training procedure is the same procedure we used for the UVCGANv2 generator pre-training.

At the second step, we trained the actual I2I translators, with the help of the provided training scripts [41]. To closely match the UVCGAN paper, we performed a hyperparameter sweep over λ_{cyc}, λ_{GP}, and γ. We explored λ_{cyc} values of $[1, 5, 10]$. For the GP pa-

rameters (λ_{GP}, γ), we explored the following configurations $[(10, 1), (1, 10), (0.1, 100), (1, 750)]$, suggested by the UVCGAN source code repository.

The best UVCGAN performance for all the high-quality translation directions (Male-to-Female, Cat-to-Dog, Wildlife-to-Dog) is achieved for the same configuration of the tested hyperparameters: $\lambda_{cyc} = 5, \lambda_{GP} = 0.1, \gamma = 100$. This configuration is used when presenting the UVCGAN results in the main part of the paper.

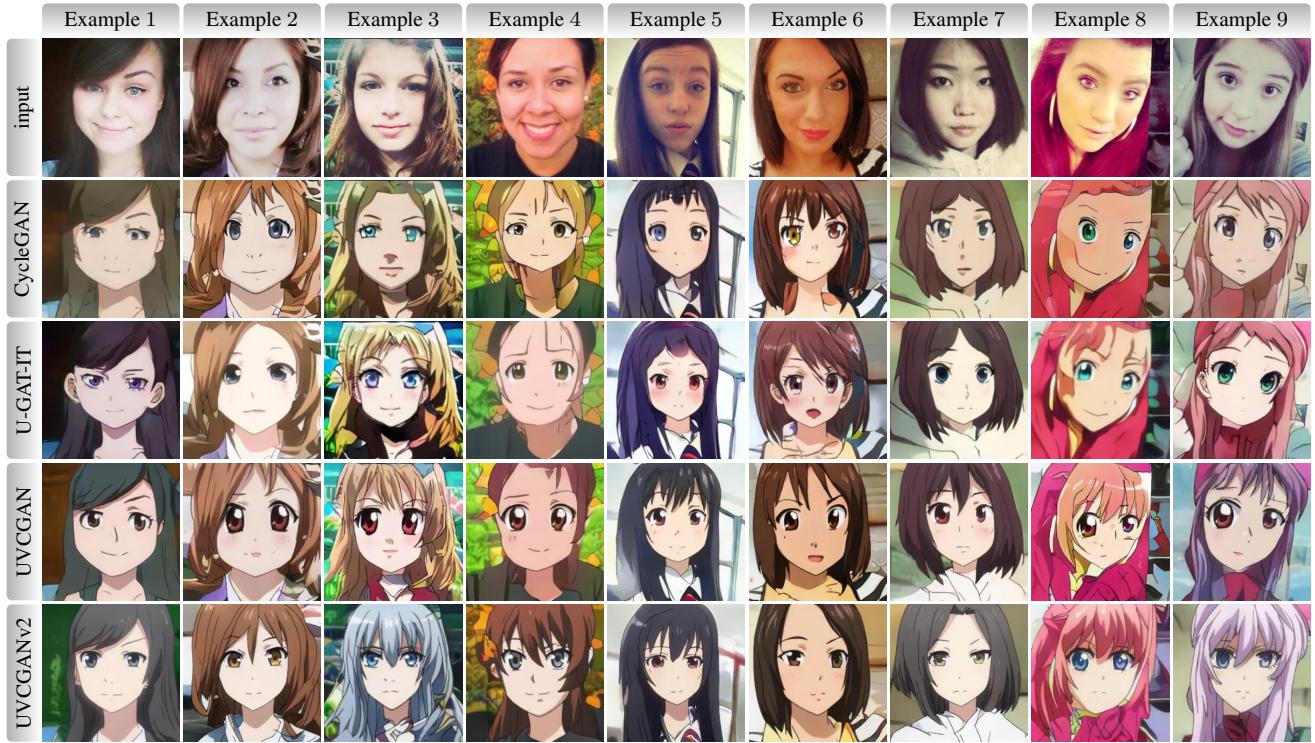Table 13. **Sample translations for Selfie-to-Anime on Anime.**

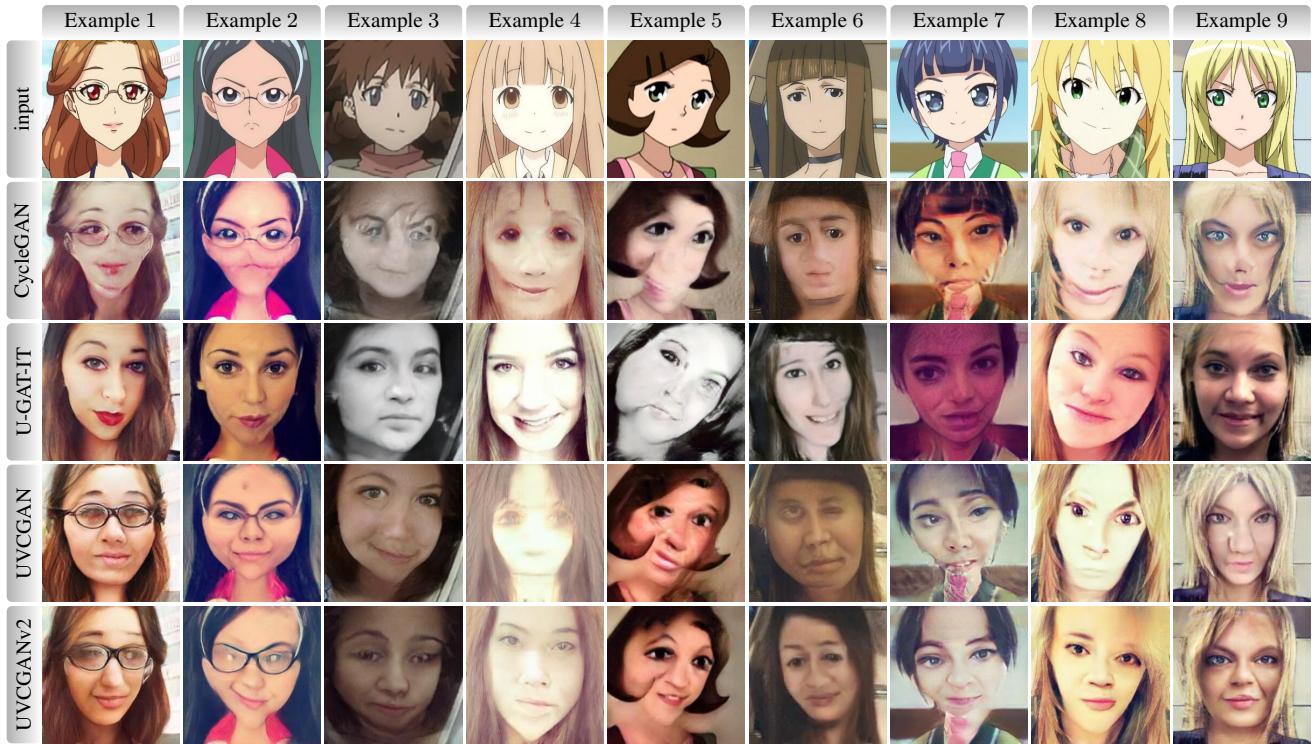Table 14. **Sample translations for Anime-to-Selfie on Anime.**

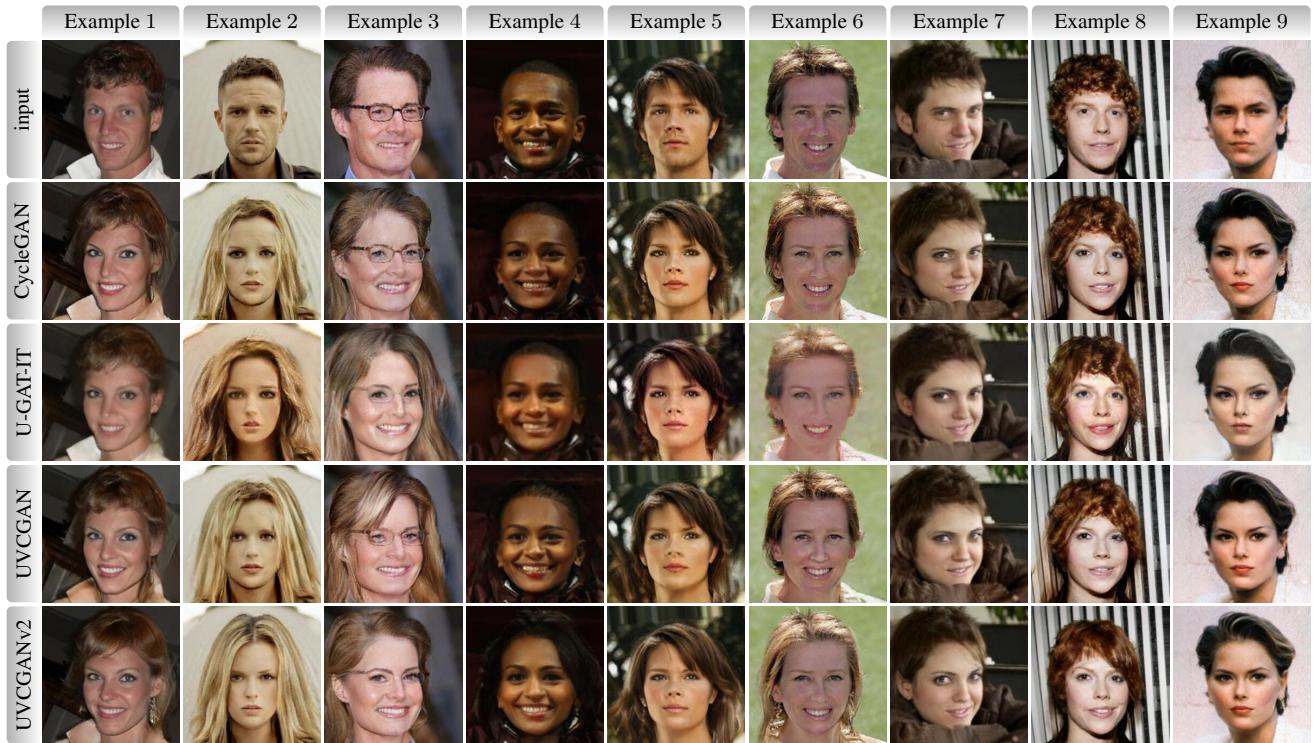Table 15. **Sample translations for Male-to-Female on CelebA.**
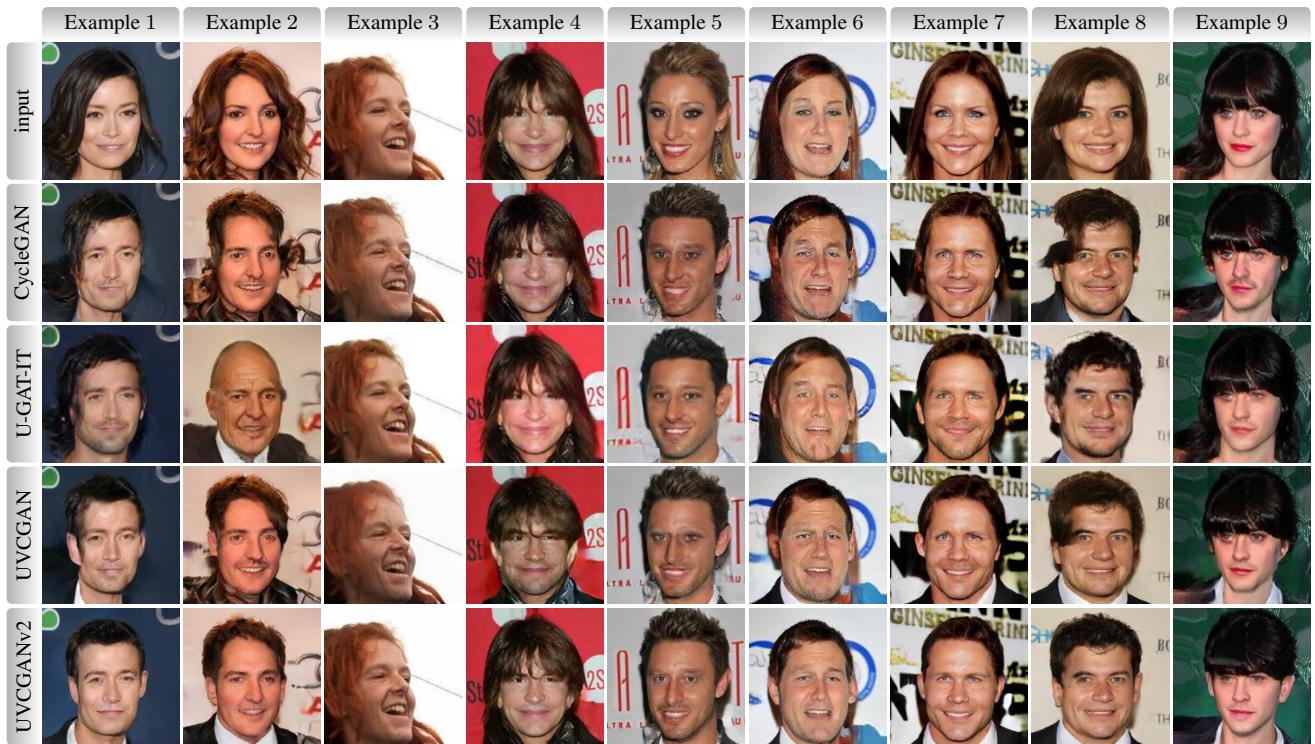

Table 16. **Sample translations for Female-to-Male on CelebA.**

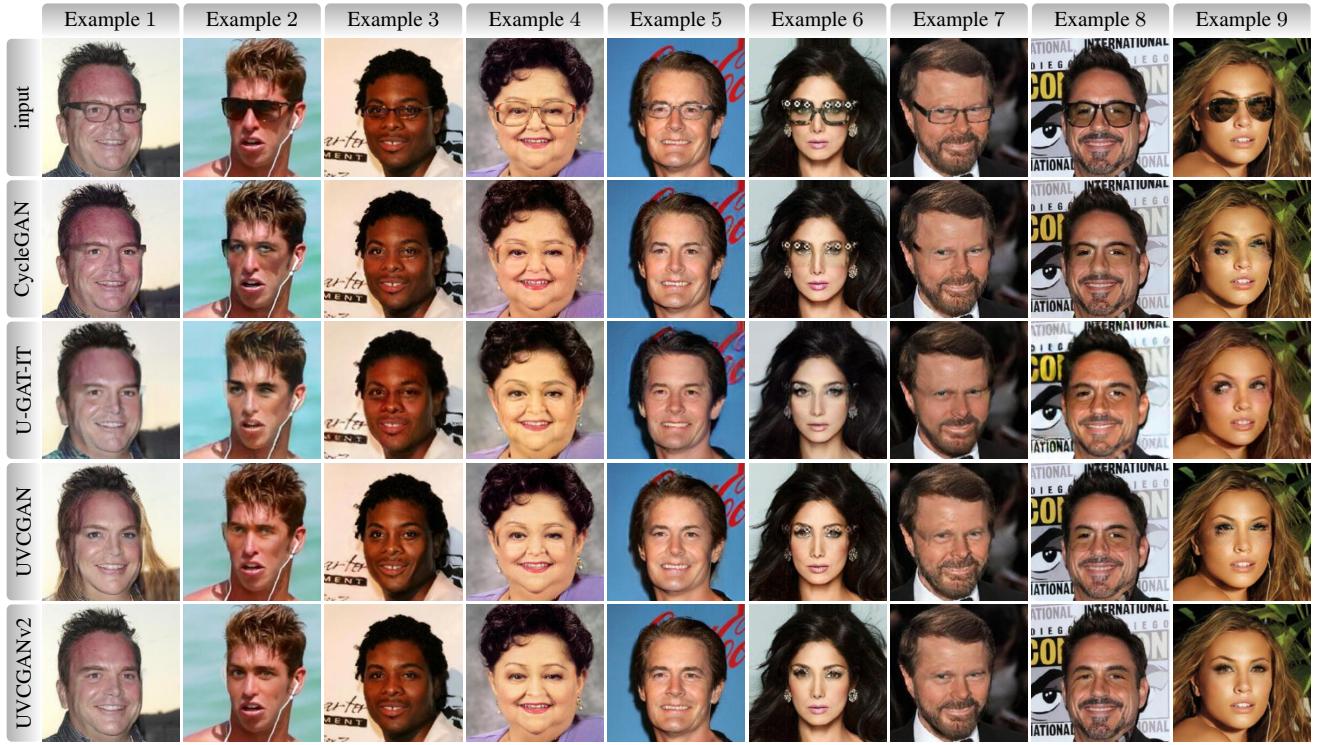Table 17. **Sample translations for Removing Glasses on CelebA.**

	Example 1	Example 2	Example 3	Example 4	Example 5	Example 6	Example 7	Example 8	Example 9
input									
CycleGAN									
U-GAT-IT									
UVCGAN									
UVCGANv2									

Table 18. **Sample translations for Adding Glasses on CelebA.**

	Example 1	Example 2	Example 3	Example 4	Example 5	Example 6	Example 7	Example 8	Example 9
input									
CycleGAN									
U-GAT-IT									
UVCGAN									
UVCGANv2									

Table 19. **Sample translations for Cat-to-Dog on AFHQ.**

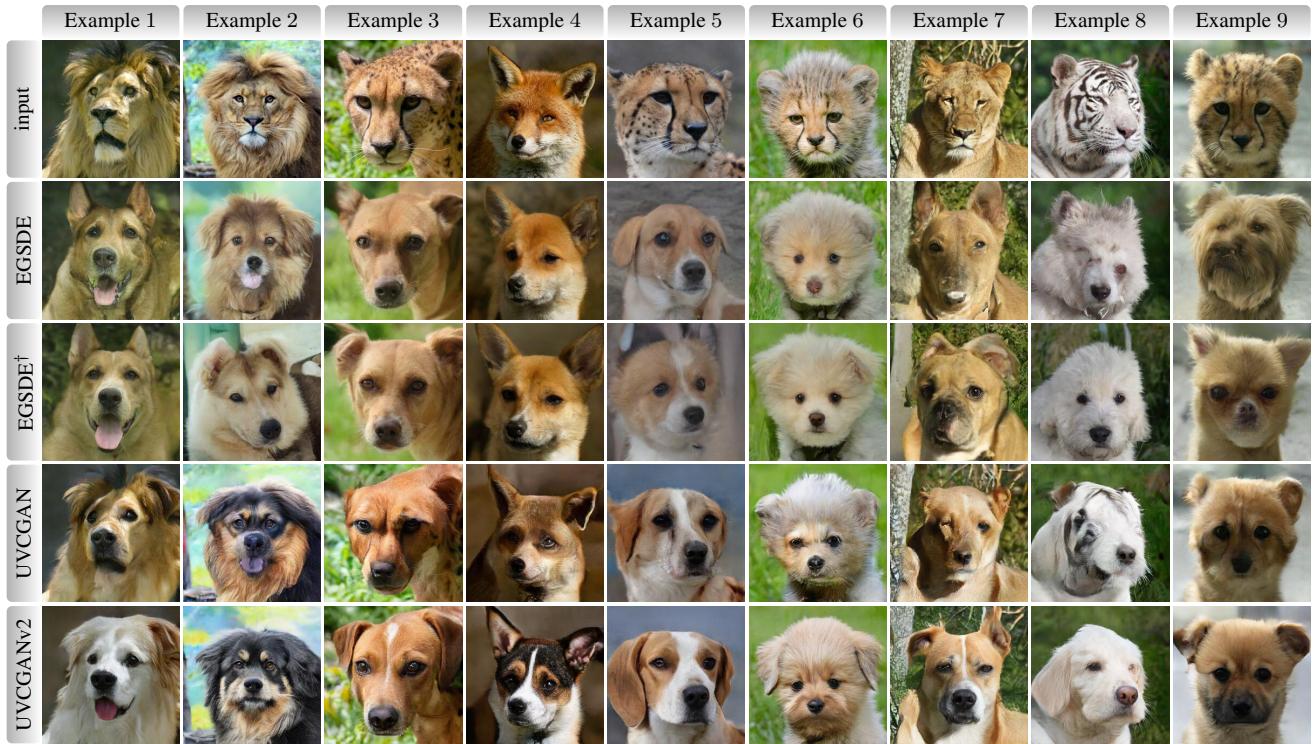Table 20. **Sample translations for Wild-to-Dog on AFHQ.**

Table 21. **Sample translations for Wild-to-Cat on AFHQ.** Since no benchmarking algorithms studied this task, we only show the input and UVCGANv2's translation.

Table 22. **Sample translations for Male-to-Female on CelebA-HQ.**