GRIT: GAN Residuals for Paired Image-to-Image Translation

Saksham Suri* Moustafa Meshry* Larry S. Davis Abhinav Shrivastava
University of Maryland, College Park

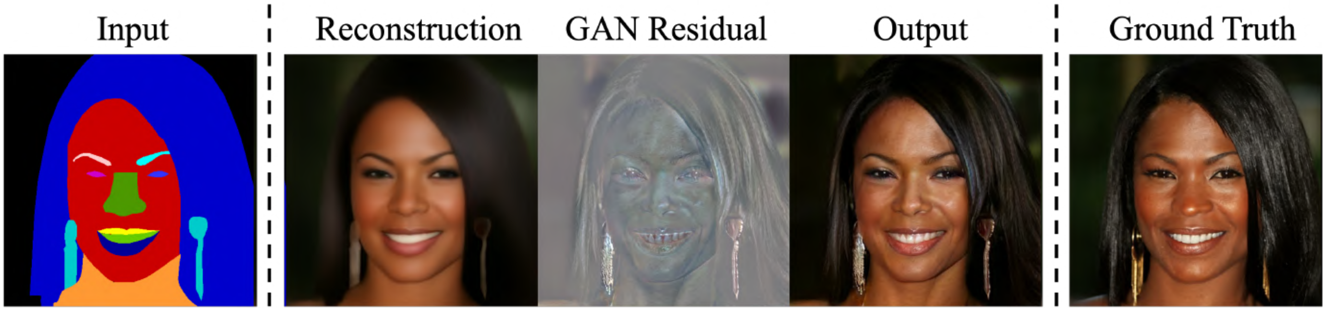

Figure 1. We decouple the optimization of reconstruction and adversarial losses by synthesizing an image as a combination of its reconstruction (low-frequency) and *GAN residual* (high-frequency) components. The GAN residual adds realistic fine details while avoiding the pixel-wise penalty imposed by reconstruction losses.

Abstract

Current Image-to-Image translation (I2I) frameworks rely heavily on reconstruction losses, where the output needs to match a given ground truth image. An adversarial loss is commonly utilized as a secondary loss term, mainly to add more realism to the output. Compared to unconditional GANs, I2I translation frameworks have more supervisory signals, but still their output shows more artifacts and does not reach the same level of realism achieved by unconditional GANs. We study the performance gap, in terms of photo-realism, between I2I translation and unconditional GAN frameworks. Based on our observations, we propose a modified architecture and training objective to address this realism gap. Our proposal relaxes the role of reconstruction losses, to act as regularizers instead of doing all the heavy lifting which is common in current I2I frameworks. Furthermore, our proposed formulation decouples the optimization of reconstruction and adversarial objectives and removes pixel-wise constraints on the final output. This allows for a set of stochastic but realistic variations of any target output image. Our project page can be accessed at cs.umd.edu/~sakshams/grit.

1. Introduction

Generative Adversarial Networks (GANs) have had a revolutionary impact on generative modeling and image synthesis. In their unconditional setting [7, 13, 15], GANs map a source distribution, typically a unit Gaussian, to a target distribution (*e.g.*, real images). At inference time, random images can be synthesized by sampling latent codes from the source distribution and passing them through a generator network. To provide user control over the synthesis process, Isola *et al.* [11] proposed a GAN-based Image-to-Image (I2I) translation framework, which conditions the synthesis process on an input image that describes certain attributes of the target output. Therefore, I2I translation learns to map images from a source domain A to a target domain B (*e.g.*, semantic maps → scenes or sketches → photo-realistic images). I2I translation has since been utilized for many problems in computer vision and graphics, such as inpainting [34], colorization [49, 51], super-resolution [20], image de-noising [4], rendering [27,28,37], and many more [5, 41, 52].

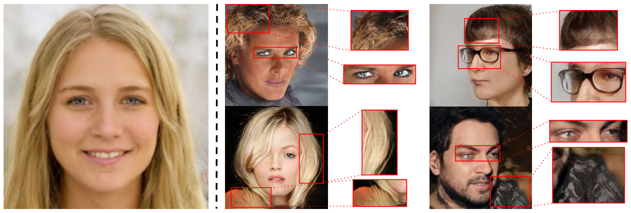

Figure 2. Comparing image realism between unconditional GANs and I2I translation. Left: Sample output from *StyleGAN* [15] at 1024 × 1024 resolution. Right: I2I translation outputs from *GauGAN* [32] at 256 × 256 resolution. Even at a lower resolution, I2I shows more noticeable artifacts compared to unconditional GANs.

[*] Equal contributors.

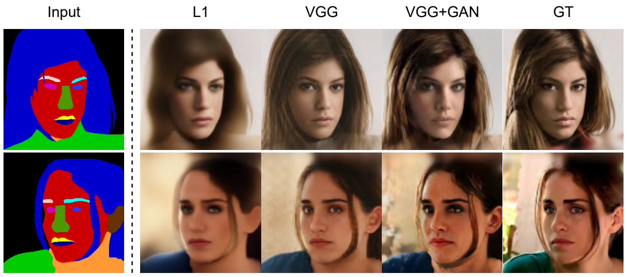

Figure 3. Comparison between different I2I training objectives: Left is the input semantic layout. The following columns show the output of networks trained with an $L1$ loss, VGG-based perceptual loss, perceptual+adversarial (GAN) losses respectively. Last column shows the corresponding ground truth image.

While unconditional GANs [14, 17] and class-conditional GANs [3] have reached unprecedented visual quality, I2I translation lags behind in quality and realism. This is despite the fact that it has more inputs and better supervision during training. For example, Figure 2 contrasts *StyleGAN* [15] (unconditional) and *GauGAN* [32] (I2I) which both came out around the same time and from the same institution. Yet, there is a clear realism gap in favor of unconditional GANs.

We are motivated by this realism gap between unconditional GANs and I2I translation. We investigate the cause of this performance gap and trace it back to the difference in the training objective between those two tasks. In unconditional GANs, the generated output is supervised only by an adversarial loss $\mathcal{L}_{\text{adv}}$, where a critic/discriminator network learns to score how realistic the output looks. On the other hand, I2I translation relies on cyclic [56] and cross-cyclic [10, 23] reconstruction losses between the generated output and available ground truth images. Thus the training objective of I2I translation optimizes a weighted combination of both an adversarial loss, $\mathcal{L}_{\text{adv}}$, and a reconstruction loss, $\mathcal{L}_{\text{rec}}$. Reconstruction losses enforce a form of pixel-wise matching between the ground-truth image I^B and the output reconstruction $\hat{I}^B$. This provides a strong supervisory signal which speeds up convergence significantly when compared to unconditional GANs. However, we show that reconstruction losses are at odds with adversarial losses, which does not lead to a sound optimization objective and causes visual artifacts in I2I outputs.

Figure 3 shows the effect of optimizing different objectives for I2I translation. Optimizing an $L1$ reconstruction alone leads to very blurry outputs. While using a VGG-based perceptual loss [12] achieves much better results, the output is not sharp and contains clear grid artifacts. Adding an adversarial loss brings the output closer to the distribution of real images, but, in many cases, artifacts can be spotted (*e.g.*, around the hair, teeth and eyes). We hypothesize that directly optimizing a reconstruction loss on the output ignores a type of multi-modality in image synthesis, which leads to visible artifacts. To motivate our hypothesis, Figure 4 shows how GANs improves realism by simulating fine details found in real images, like local variations or noise patterns found in the texture of real materials. There are infinite realizations of such noise patterns that add realism to the output (*e.g.*, skin freckles, pores and wrinkles, and linings of hair strands). However, applying a reconstruction loss ($\mathcal{L}_{\text{rec}}$) in traditional I2I frameworks penalizes all these local variations and promotes a uni-modal solution where the generated image matches the ground truth down to the pixel level. This leads to smoothed outputs and other noticeable artifacts that do not show in the unconditional GAN setup where no reconstruction loss is applied.

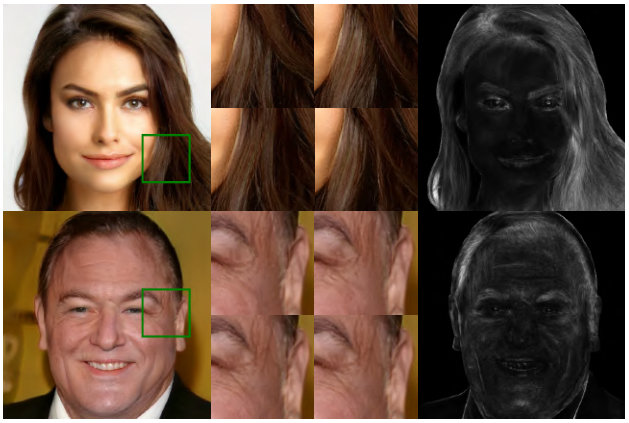

Figure 4. Examples of multi-modal outputs (generated by our method) with local stochastic variations that add realism and satisfy the GAN objective. Applying reconstruction losses in traditional I2I frameworks ignores this type of multi-modality and penalizes such variations, which misleads I2I training.

In this work, we address this problem and propose a modified architecture and training objective that relaxes the role of reconstruction losses to act as regularizers instead of doing all the heavy lifting which is common in current I2I translation frameworks. Our formulation decouples the optimization of adversarial and reconstruction losses. This enables our network to hallucinate local variations to add realism to the output while avoiding being penalized by reconstruction losses. Although we investigate our proposal in a paired I2I setting, the idea can be extended to unpaired I2I.

We summarize our contributions as follow:

- We study the realism gap between unconditional GANs and paired I2I translation, and shed light on an important multi-modal aspect of image synthesis that we denote as *local spatial variations*, which is overlooked and rather penalized in traditional I2I translation formulation.
- Through the proposed approach, we use GAN Residuals for Image-to-Image Translation (GRIT), and take the first step towards addressing the multi-modal nature of local

spatial variations in I2I translation. We utilize a modified architecture and training objective that models and encourages such multi-modality.

- We provide quantitative and visual evidence on the effectiveness of modeling local spatial variations in paired I2I translation, and show that our proposed method improves upon strong baselines.

2. Related work

Since the onset of the GAN era with the seminal work of Goodfellow *et al.* [7] there have been multiple works [3, 13, 16, 30] to improve the synthesis quality and resolution of images. Karras *et al.* [17] improved the quality of [16] by introducing better normalization and regularization and in the process reduced image artifacts and made inversion easier. [14] went a step further to reduce aliasing and also proposed an approach which made representations equivariant to rotation and translation. While these works try to learn the manifold of training data to generate samples on it, there is another synthesis task of Image-to-Image (I2I) translation which has been vastly explored and involves translating images from one domain to another. I2I translation can be broadly grouped into two regimes based on the type of training data, unpaired data or paired data.

Unpaired Image-to-Image Translation utilizes unpaired training data which does not have pixel level correspondence between the domains. CycleGAN [55], DualGAN [46] and DiscoGAN [18] proposed one of the first and most commonly used approaches for this involving a cyclic loss to impose consistency between forward and backward translation for the same image. UNIT [25] and SCAN [24] also utilize the cyclic loss but introduce a shared latent space and multistage coarse to fine training respectively. TransGaGa [44] extends the cyclic loss to large domain gaps by disentangling features into appearance and geometry latent space. On the uni-directional (non-cyclic loss) end, approaches like DistanceGAN [2] train by maintaining distance between pairs of samples. GcGAN [6] enforces constraints on geometric transformations preserving image semantics. CUT [31] proposes a multi-layer patch based contrastive learning approach while MUNIT [10] and DRIT [22] disentangle representations into style and content. [33] also uses disentanglement but into texture and structure. More recently MSPC [45] proposes a maximum spatial perturbation consistency based regularization.

Paired Image-to-Image Translation uses paired data making it possible to enforce pixel level correspondence. One of the first works in this direction, Pix2Pix [11] proposed an $L1$ reconstruction loss with a patch discriminator. Later, Pix2PixHD [39] improved it with higher resolution generation using coarse to fine generator and multi-scale discriminator. DNI [42] looks at decoupling reconstruction and GAN losses but they end up training two separate models with differing objectives and then interpolating between them by performing a weighted summation to get a balance between both tasks. They also assume some level of correlation between the parameters of the two networks for this to work and lack quantitative evidence for the efficacy of their approach. SPADE [32] proposed spatially-adaptive normalization layer as vanilla normalization washes away semantic information. SEAN [57] further proposed semantic region-adaptive normalization layer to control style of each semantic region separately. CoCosNet [48] jointly learns the cross domain correspondence and image translation, where both tasks facilitate each other and thus can be learned with weak supervision. Later CoCosNetv2 [53] mitigated the quadratic complexity issue in CoCosNet and enabled high-resolution correspondence using PatchMatch [1]. Recently DINO [38] proposed an energy based cyclic framework to utilize the conditional input. While MoNCE [47], presents a re-weighted patch based constrastive learning framework. Unlike these works in our approach we disentangle the reconstruction and adversarial (GAN) loss. Additionally, we also propose architecture modifications which enable us to perform this disentanglement by separating the reconstruction supervised output and the residual and in the process make better use of per-pixel spatial noise to learn more realistic and diverse I2I translations.

3. Approach

The goal of reconstruction losses is to guide a generated output $\hat{I}^B$ to resemble a target ground truth image I^B. While this is a desired behavior for I2I translation, a negative side effect is that a reconstruction loss will also penalize high frequency deviations between $\hat{I}^B$ and I^B. Therefore, this formulation ignores the multi-modal nature of synthesizing fine-grain texture patterns, where there are infinitely many realizations of local high-frequency details (*e.g.*, skin texture or the location of hair strands as shown in Figure 4). Penalizing such local variations and promoting a uni-modal solution thus causes artifacts and contributes to the realism gap between unconditional GANs and I2I translation.

Through our approach GRIT, we make a first step towards addressing this overlooked multi-modal aspect of image synthesis, and propose to decouple the optimization of reconstruction and adversarial losses.

We present our formulation in Section 3.1 and associated changes to the loss function in Section 3.2. Section 3.3 discusses how to explicitly model multi-modal local variations for paired I2I translation.

3.1. Formulation

We propose to generate an image $\hat{I}^B$ as the composition of two components: a reconstruction component $\hat{I}^B_{\text{rec}}$, and

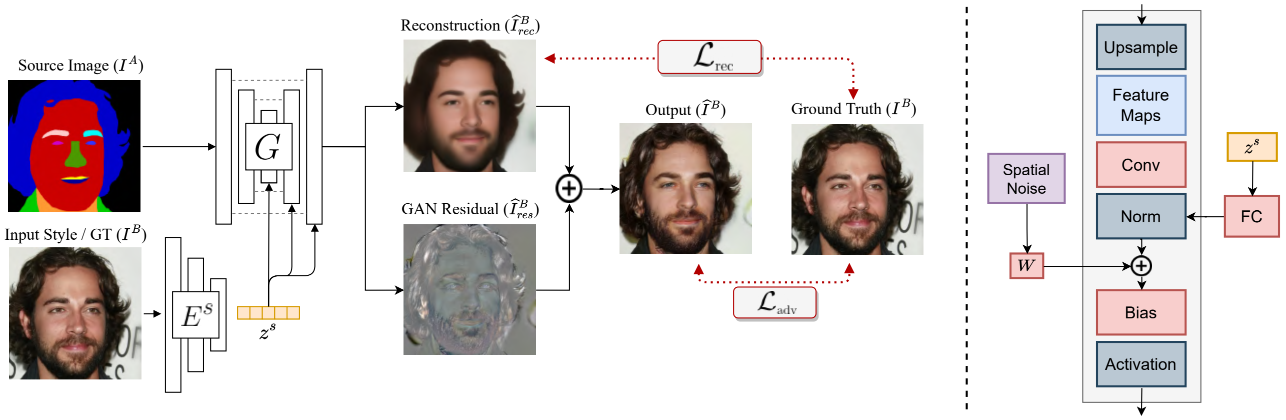

Figure 5. **Left:** : Overview of GRIT. Our network generates the output as the composition of a reconstruction component $\hat{I}^B_{\text{rec}}$ and a *GAN-residual* component $\hat{I}^B_{\text{res}}$. An $L1$ reconstruction loss is applied only to the reconstruction component, while the GAN residual is supervised only through an adversarial loss $\mathcal{L}_{\text{adv}}$. **Right:** The generator's upsampling block. We feed the encoded style latent z^s through *AdaIN* layers, and also add random spatial noise maps controlled by learnable weights W to the feature maps.

an adversarial *GAN-residual* component $\hat{I}^B_{\text{res}}$. During training, this decoupling of $\hat{I}^B_{\text{rec}}$ and $\hat{I}^B_{\text{res}}$ allows the reconstruction component $\hat{I}^B_{\text{rec}}$ to focus on reconstructing low-frequency details of the target real image I^B, while the *GAN-residual* component $\hat{I}^B_{\text{res}}$ hallucinates high-frequency details that add realism to the synthesized image $\hat{I}^B$. The final output is generated as:

$$\hat{I}^B = \mathcal{C}(\hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}}); \qquad \hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}} = G(I^A, z^s) \qquad (1)$$

where $G(.,.)$ is a generator network that maps an input image I^A along with a style latent code z^s to its low- and high-frequency components $\hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}}$, and $\mathcal{C}(.,.)$ is a composition operator that combines both output components. We implement $\mathcal{C}$ as a simple addition. We also investigated implementing it as a small CNN head that fuses $\hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}}$ but found that simply adding the two images works better in our case and is more stable to train.

Figure 5 gives an overview of our architecture. We implement the generator network as a U-Net [35] architecture that consists of a content encoder E^C and a decoder D. The decoder network outputs both $\hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}}$. We further discuss the decoder architecture in Section 3.3.

To model style multi-modality, we follow the literature [29,32,56] by utilizing a style encoder E^S that learns to capture the style of an input image into a latent style code z^s, which is fed to the generator G via *AdaIN* layers [9] to specify the style of the output $\hat{I}^B$. Next, we discuss our modification to the loss function to encourage the decomposition of $\hat{I}^B$ into its reconstruction and GAN-residual components.

3.2. Loss function

Standard loss function of GAN-based I2I translation networks consists of a weighted sum of a pixel-wise reconstruction loss $\mathcal{L}_{\text{rec}}$ and a discriminator-based adversarial loss $\mathcal{L}_{\text{adv}}$.

Minimizing this loss does not take into account possible local variations, as it promotes pixel-wise matching between the output $\hat{I}^B$ and the ground truth I^B, and thus only accepts one solution and penalizes any high-frequency variations. To allow local variations, we aim to only reconstruct the low-frequency components of a ground truth image I^B, where low-frequency components capture the *general* content and style of the target output. On the other hand, we want the generator to have the freedom to add fine-grain details, represented by high-frequency components, making the output photo-realistic.

We achieve this by modifying the loss to apply the reconstruction loss $\mathcal{L}_{\text{rec}}$ only to the reconstruction component $\hat{I}^B_{\text{rec}}$, while the adversarial loss $\mathcal{L}_{\text{adv}}$ is applied to the final output $\hat{I}^B = \mathcal{C}(\hat{I}^B_{\text{rec}}, \hat{I}^B_{\text{res}})$. Thus, our modified training objective is given by:

$$\min \mathcal{L}(I^B, \hat{I}^B, \hat{I}^B_{\text{rec}}) = \mathcal{L}_{\text{adv}}(\hat{I}^B, I^B) + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}(\hat{I}^B_{\text{rec}}, I^B) \qquad (2)$$

With such modification, the reconstruction loss $\mathcal{L}_{\text{rec}}$ does not backpropagate into the GAN-residual component $\hat{I}^B_{\text{res}}$, and $\hat{I}^B_{\text{res}}$ therefore has the freedom to hallucinate high-frequency details that add realism to generated images without being constrained to match pixel-level details of ground truth images at training time. While the proposed loss function (Eqn. 2) allows high-frequency deviations between I^B and $\hat{I}^B$, this by itself does not encourage multi-modal synthesis of local texture and other high-frequency details. In the next section, we discuss how to explicitly model the local-variations multi-modality into our network.

3.3. Multi-modal outputs

At training time, I2I translation networks peek at the target ground truth image I^B and encodes it into a flattened

Table 1. Comparison with baselines at 256 × 256 resolution.

Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pix2PixHD [40]	24.78	17.53	0.515	0.256	45.97
SPADE [32]	28.69	16.21	0.485	0.283	26.06
DINO [38]	51.84	12.13	0.401	0.369	37.24
MoNCE [47]	64.26	10.32	0.357	0.380	34.27
Ours	**18.34**	**19.54**	**0.531**	**0.245**	**17.04**

style latent code z^s. However, due to the lossy nature of such compression, it is impossible for the decoder to recover pixel-level spatial information (*e.g*., location of hair strands) to reconstruct I^B. Driven by the adversarial loss, the decoder hallicinates spatial patterns to bring synthesized images closer to the manifold of real images. This requires the decoder to devise a way to generate spatially-varying pseudo-random numbers from the input flattened latent. This challenge was first raised in StyleGAN [15], where they showed that this is inefficient and consumes much of the network capacity. To address this limitation, Karras *et al.* [15] proposed to add per-pixel noise maps within each upsampling block in the decoder to encourage synthesizing local variations of spatial patterns.

The use of spatial noise maps however did not transfer to the I2I translation literature. This is because, unlike unconditional GANs, the application of reconstruction losses counteracts the added spatial noise by suppressing it, leading to a uni-modal output. On the other hand, decomposing the synthesis into its reconstruction $\hat{I}^B_{\text{rec}}$ and GAN-residual $\hat{I}^B_{\text{res}}$ components allows for naturally adapting spatial noise maps to I2I translation by modeling local stochastic variations in the GAN-residual component. This bridges the gap between unconditional GANs and I2I translation since $\hat{I}^B_{\text{res}}$ is not affected by the reconstruction loss, and can therefore fully utilize the added spatial noise. Adding spatial noise maps models the local variation multi-modality, and enables generating multi-modal output for the same target image by sampling random noise maps.

4. Experimental evaluation

Implementation details. In the interest of space we provide details about network architecture and training hyper-parameters in the supplementary.

Dataset. We perform our main evaluation on the CelebAMask-HQ dataset [21]. The dataset contains $30,000$ high resolution face images along with their corresponding segmentation masks which contain 19 semantic labels and are at a 512 × 512 resolution. We use the standard train and test splits provided by Liu *et al*. [26]. We also show results on Edges2Handbags [54] which contains 137K Amazon Handbag images and edge maps. All images and edge maps are at 256 × 256 resolution. Unless stated otherwise, all experiments and analysis are performed on the CelebAMask-HQ dataset [21].

Table 2. Comparison on Edges2Handbags at 256 × 256 image resolution.

Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pix2PixHD [40]	19.28	18.05	0.70	0.20	59.53
SPADE [32]	21.91	16.71	0.66	0.25	75.49
Ours	**12.61**	**20.68**	**0.73**	**0.18**	**57.79**

Baselines. We compare our method with the following approaches: Pix2PixHD [39], SPADE (also called GauGAN) [32], DINO [38] and MoNCE [47]. We train Pix2PixHD [39] and SPADE [32] using their official released code. For Pix2PixHD, we enable the option to train a semantic-specific style encoder, which computes separate style codes per semantic label. We use the outputs provided by the authors for DINO [38], and use the released pre-trained model of MoNCE [47] and follow the authors' instructions to generate test results on the CelebAMask-HQ dataset. Since MoNCE and DINO are trained at 256 × 256, we train our method as well as Pix2PixHD and SPADE at 256 resolution for fair comparison. Additionally, we also train our method, Pix2PixHD and SPADE at 512 × 512 resolution to evaluate and compare results at high resolution.

Metrics. We evaluate using the following metrics:

- Standard reconstruction metrics such as $L1$, Peak Signal to Noise Ratio (PSNR), and structural similarity (SSIM) [43] between the output and ground truth.
- LPIPS [50] which measures the perceptual similarity between the output and ground truth using AlexNet features.
- Frechet-Inception Distance (FID) [8] which is used to measure the perceptual quality and realism of the output.

4.1. Quantitative Comparison

We provide quantitative comparison with the baselines in Table 1 and 2 for the CelebAMask-HQ and Edges2handbags datasets respectively which are commonly used in paired I2I literature. For CelebAMask-HQ dataset, we observe that Pix2PixHD performs much better than SPADE on reconstruction metrics like L1, PSNR, SSIM and LPIPS for both the datasets. This is because Pix2PixHD uses a powerful semantic-specific style encoder that encodes a separate style code per each semantic label and is therefore able to match the ground truth style more accurately. On the other hand, SPADE uses a VAE-based encoder [19] which adds robustness to noise in the style latent space, but at the expense of faithful reconstruction of the ground truth style. SPADE however maintains good realism, and thus performs much better than Pix2PixHD on the FID metric. While DINO [38] and MoNCE [47] are more recent baselines, we observe they fall short in comparison

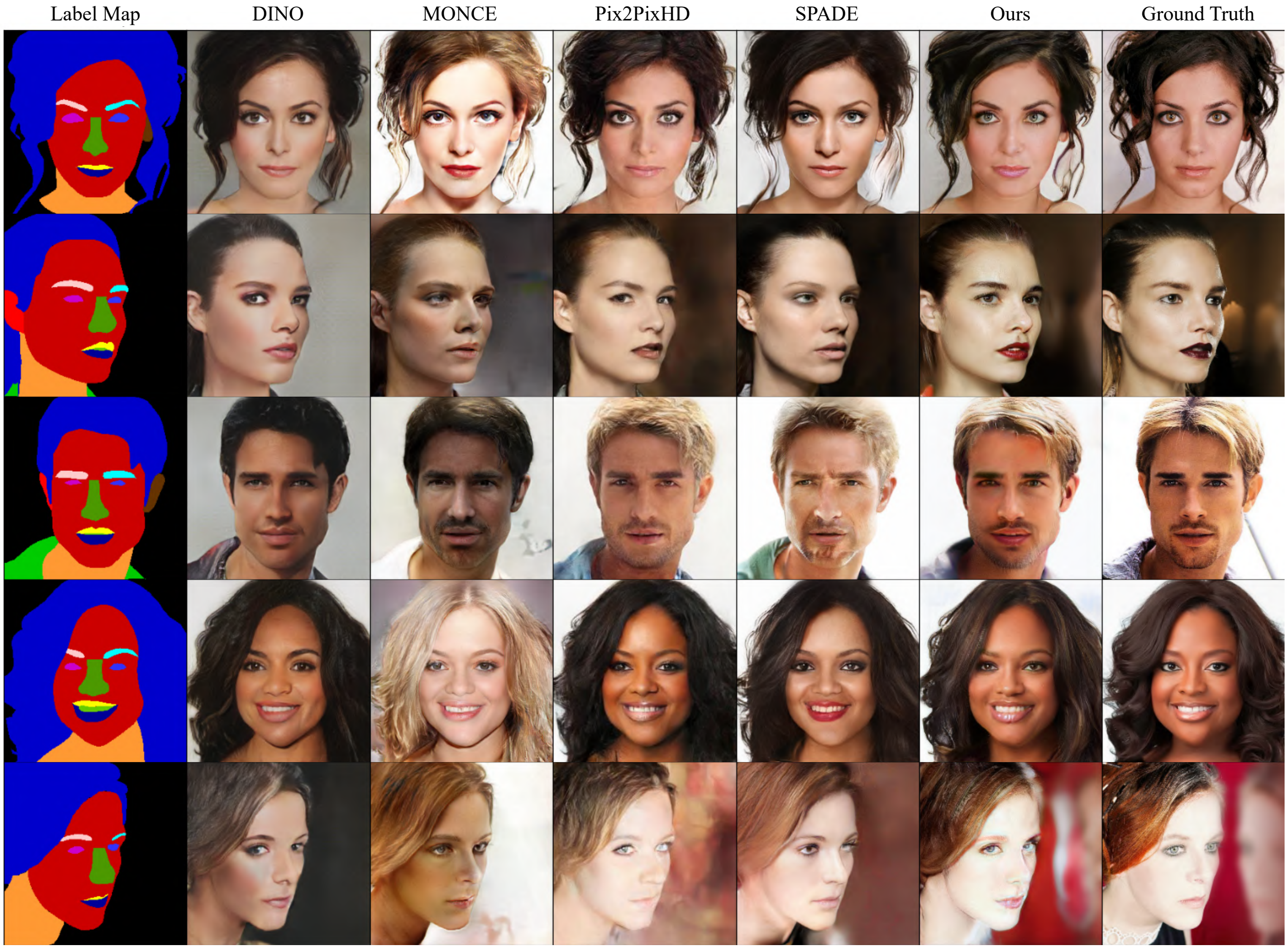

Figure 6. Qualitative comparison on CelebAMask-HQ dataset with DINO [38], MoNCE [47], Pix2PixHD [40] and GauGAN/SPADE [32].

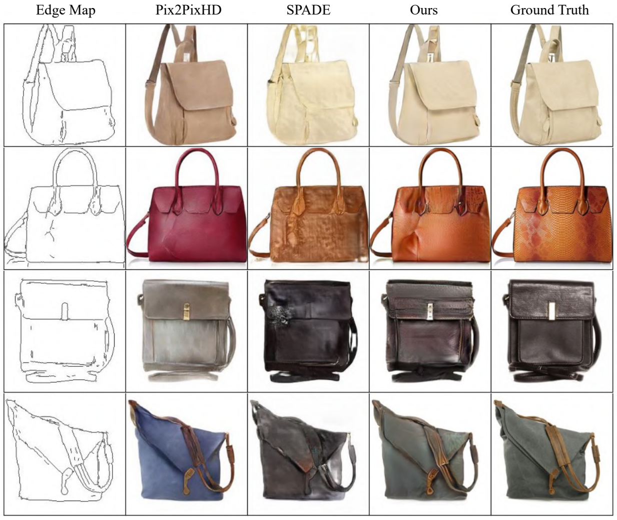

Figure 7. Qualitative comparison on Edges2Handbags dataset with Pix2PixHD [40] and SPADE [32].

with Pix2PixHD and SPADE. Finally, our decoupled optimization of reconstruction and adversarial losses achieves better reconstruction error, as well as better realism (FID) score compared to the baselines. Our reconstruction component $\hat{I}^B_{\text{rec}}$ focuses on reconstructing low-frequency details to match the general color and structure of the ground truth. Matching low-frequency components has a direct impact on reconstruction metrics, especially L1 and PSNR. Additionally, unlike the baselines, our GAN residual component $\hat{I}^B_{\text{res}}$ is not constrained by reconstruction losses. And so, it has the freedom to add high-frequency details that improves the output realism, which leads to a better FID score. Similar trends hold for the Edges2Handbags dataset with Pix2PixHD performing better than SPADE on all metrics including FID. This is because SPADE is designed for dense spatial inputs, *e.g.* semantic maps, not sparse edge maps as in the case of Edges2Handbags.

While many baselines are trained at a 256×256 resolution, we also inspect our performance at a higher resolution of 512×512 on the CelebAMask-HQ dataset. To provide comparative evaluation at this resolution, we choose the Pix2PixHD and SPADE methods which are the top per-

Table 3. Comparison with baselines at resolution of 512 × 512.

Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pix2PixHD [40]	24.80	17.40	0.534	0.354	24.79
SPADE [32]	31.44	15.53	0.490	0.389	20.80
Ours	**19.02**	**19.36**	**0.555**	**0.333**	**16.91**

forming methods at 256 × 256 resolution, and retrain them at a 512 resolution. Table 3 shows that the proposed method consistently shows similar trends of improvement over the baselines across all metrics. We qualitatively show examples for this resolution in the supplemental.

4.2. Qualitative evaluation

In this section we qualitatively analyze and compare synthesized results between our method and the baselines. We also look at various aspects of our approach through visual results to understand different components better.

4.2.1 Comparison with Baselines

Figure 6 shows qualitative comparison with the baselines on the CelebAMask-HQ dataset. Our method clearly improves over the baselines in terms of both realism, as well as matching the ground truth style. We note that our results could show some style deviations from the ground truth style (*e.g.*, lip color in the second row), we are still noticeably better than the baselines. We observe that the added GAN residual can sometimes cause such slight deviation from the reconstructed color, since it is not constrained by the reconstruction loss. While DINO captures the structure well, it loses out on realism and on matching colors and textures to the ground truth image. MoNCE shows more details due its patch based nature during training, but again is not able to faithfully capture the style and structure well. Pix2PixHD and SPADE both generate reasonable results, but we observe that SPADE results look more realistic, although not faithfully matching the ground truth style. Our output on the other hand generates high quality and realistic samples while making sensible light deviations which capture the true nature of real world data.

We show results on Edges2Handbags dataset in Figure 7 where we observe similar trends as the Pix2PixHD output looks better than SPADE and has fewer artifacts while ours looks the most faithful. As can be seen from the figure our method generates the color, texture and structure better compared to the other approaches.

4.2.2 Standard Deviation of Spatial Noise

In our approach we utilize spatial noise by adding it to the feature maps at each upsampling block. This along with the decoupled objective lets the network learn to generate variations of local information which preserves the structure and content but introduces diversity in the generated samples. Here we analyze the variations generated for multiple subjects over 20 different spatial noise samples for each of them to understand the stochasticity better. Figure 8 shows pixel-wise standard deviation over the different translation results generated by varying the spatial noise on CelebAMask-HQ dataset. It can be see that highest deviation occurs in regions corresponding to hair, around eyes, lips and nose. These regions can be considered high-frequency locations as they usually contain multiple edges and have the most variations. By visualizing the standard deviation we are able to verify that the network is able to understand and model these regions better and generate sensible variations.

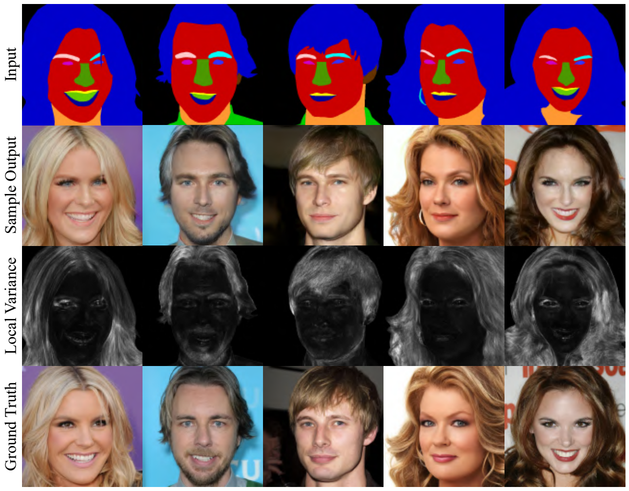

Figure 8. Examples of local stochastic variations. Top to bottom rows represent the input image, one sample output, standard deviation of each pixel over 20 different outputs for the same sample, and ground truth image respectively.

4.3. Ablation

We perform ablation our approach to show the effect of sequentially introducing each component using the CelebAMask-HQ dataset. These results are highlighted in Table 4. The first row shows the performance of a vanilla I2I framework which utilizes a U-Net [35] based generator with a reconstruction and GAN loss on the output. The second row corresponds to introducing spatial noise which lets the model learn to generate local variations. It should be noted that introducing spatial noise on its own does not realize its full potential as the reconstruction loss can fight back and teach the network to ignore it in order to improve on the pixel-wise reconstruction loss. This is where the role of residuals comes in, which can be seen in the third row. Introducing the GAN residuals along with the spatial noise gives a considerable boost in performance, as while the reconstruction losses supervise the reconstructed output, the GAN loss supervises the combined output and lets the network learn residuals which can better capture details in the

Table 4. Ablation of different components of our approach.

Method	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
U-Net [35]	21.68	18.25	0.504	0.222	20.24
+ spatial noise	20.93	18.55	0.520	**0.219**	21.41
+ GAN residuals (**ours**)	**18.34**	**19.54**	**0.531**	0.245	**17.04**

image. We also compare the VGG and L1 loss as possible choices for supervising the reconstruction and show results in the supplemental supporting our choice for L1.

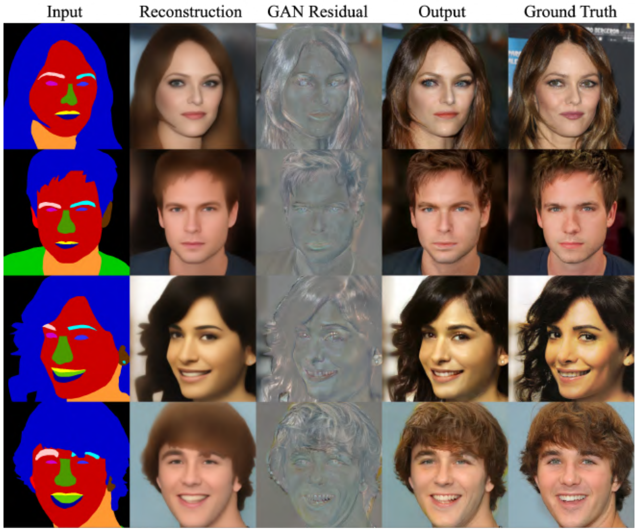

Figure 9. Examples of the different outputs of our method along with the input label map and ground truth image.

4.4. Understanding the GAN Residuals

While Figure 8 shows examples of the different outputs generated by our network, namely, the reconstructed and residual images followed by the final combined image. Here we try to understand what kind of information these images hold. As can be seen from Figure 9 the reconstructed image encodes most of the structure and content of the image. It looks like a low-frequency and smooth image while the the residual seems to contain a lot of high-frequency information around the hair, eyes, beard, lips etc. where a lot of edges and variation occur. As can be seen in the combined output, adding these two gives a realistic image which resembles the ground truth.

We refer to the GAN residuals being high-frequency by capturing local variations. Here we verify this by computing the frequency spectrum of the images in a similar manner as Schwarz *et al.* [36] who use azimuthal averaging over the spectrum in normalized polar coordinates. In Figure 10 we show the average over all the synthesized outputs corresponding to the test set for CelebAMask-HQ dataset. It can be seen how the the reconstruction (orange) encodes higher magnitude for the low-frequencies with a complete cutoff at mid-to-late frequencies. On the other hand the residual (blue), encodes more of the high-frequency information. Combining both of them (green) is much closer to the frequency spectrum of the ground truth (red).

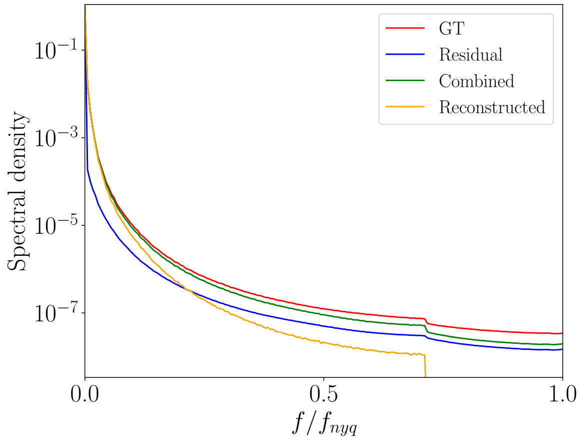

Figure 10. We visualize the frequency spectrum and highlight that the reconstructed image contains higher magnitude of low-frequency information while the residual captures the high-frequency more. By combining these, the resulting image has a spectrum closer to the ground truth image. The y-axis denotes the spectral density which is measures the magnitude of a particular frequency while the x-axis corresponds to the frequency relative to the maximum frequency corresponding to f_{nyq}.

5. Conclusion

We propose a novel approach for paired image-to-Image translation by highlighting the disconnect between the reconstruction and adversarial losses which are at odds with each other. Based on this insight we decouple the reconstruction and adversarial loss in the proposed approach which enable it to have the freedom to learn local variations better and generate more realistic translations. Through both quantitative and qualitative results we highlight the efficacy of the proposed approach and achieve state-of-the-art performance on paired I2I task on both CelebAMask-HQ and Edges2Handbags datasets. We show results and compare at both 256 × 256 and 512 × 512 resolutions which shows that the proposed method can generate higher resolution images too. We also analyze the diversity in image synthesis that our method introduces using the spatial noise and highlight its relation to high-frequency. Further, we analyze the residuals and the reconstructed output and visually show the importance of having a combination of these to give the final output along with their frequency analysis. Although we investigated our proposal in a paired I2I setting, the idea can be extended to unpaired I2I.

Acknowledgements. This project was partially funded by the DARPA SemaFor (HR001119S0085) and DARPA SAIL-ON (W911NF2020009) programs.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 3

[2] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *Int. Conf. Learn. Represent.*, 2019. 2, 3

[4] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3155–3164, 2018. 1

[5] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Int. Conf. Comput. Vis.*, 2017. 1

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 3

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, pages 2672–2680, 2014. 1, 3

[8] 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. 5

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Int. Conf. Comput. Vis.*, pages 1501–1510, 2017. 4

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Eur. Conf. Comput. Vis.*, 2018. 2, 3

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 3

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, 2016. 2

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018. 1, 3

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2, 3

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 5

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. 2, 3

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 3

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Int. Conf. Learn. Represent.*, 2014. 5

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[21] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Eur. Conf. Comput. Vis.*, 2018. 3

[23] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *Int. J. Comput. Vis.*, pages 1–16, 2020. 2

[24] Minjun Li, Haozhi Huang, Lin Ma, Wei Liu, Tong Zhang, and Yugang Jiang. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 184–199, 2018. 3

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Adv. Neural Inform. Process. Syst.*, 2017. 3

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, December 2015. 5

[27] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. LookinGood: Enhancing performance capture with real-time neural re-rendering. In *Proc. SIGGRAPH Asia*, 2018. 1

[28] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[29] Moustafa Meshry, Yixuan Ren, Larry S Davis, and Abhinav Shrivastava. Step: Style-based encoder pre-training

for multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3712–3721, 2021. 4

[30] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 3

[31] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 3

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 3, 4, 5, 6, 7

[33] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020. 3

[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015. 4, 7, 8

[36] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021. 8

[37] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 2019. 1

[38] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Dino: A conditional energy-based gan for domain translation. *Int. Conf. Learn. Represent.*, 2021. 3, 5, 6

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3, 5

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 5, 6, 7

[41] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *Eur. Conf. Comput. Vis.*, 2016. 1

[42] Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. Deep network interpolation for continuous imagery effect transition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2019. 3

[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] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8012–8021, 2019. 3

[45] Yanwu Xu, Shaoan Xie, Wenhao Wu, Kun Zhang, Mingming Gong, and Kayhan Batmanghelich. Maximum spatial perturbation consistency for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18311–18320, 2022. 3

[46] 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. 3

[47] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18280–18290, 2022. 3, 5, 6

[48] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 3

[49] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Eur. Conf. Comput. Vis.*, pages 649–666. Springer, 2016. 1

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 5

[51] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Trans. Graph.*, 36(4):119, 2017. 1

[52] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[53] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 3

[54] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016. 5

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, 2017. 3

[56] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. To-

ward multimodal image-to-image translation. In *Adv. Neural Inform. Process. Syst.*, 2017. 2, 4

[57] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5104–5113, 2020. 3