Unsupervised Image-to-Image Translation with Generative Prior

Shuai Yang   Liming Jiang   Ziwei Liu   Chen Change Loy ✉
S-Lab, Nanyang Technological University
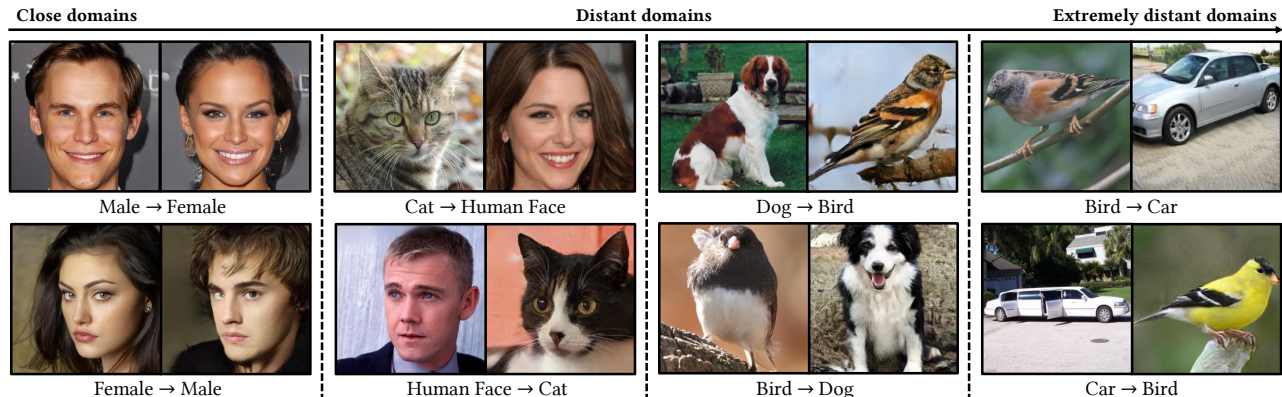{shuai.yang, liming002, ziwei.liu, ccloy}@ntu.edu.sg

Figure 1. We propose a versatile unsupervised image translation framework with generative prior that supports various translations from close domains to distant domains with drastic shape and appearance discrepancies. Each group shows *(left)* the input and *(right)* our result.

Abstract

*Unsupervised image-to-image translation aims to learn the translation between two visual domains without paired data. Despite the recent progress in image translation models, it remains challenging to build mappings between complex domains with drastic visual discrepancies. In this work, we present a novel framework, Generative Prior-guided UNsupervised Image-to-image Translation (**GP-UNIT**), to improve the overall quality and applicability of the translation algorithm. Our key insight is to leverage the generative prior from pre-trained class-conditional GANs (e.g., BigGAN) to learn rich content correspondences across various domains. We propose a novel coarse-to-fine scheme: we first distill the generative prior to capture a robust coarse-level content representation that can link objects at an abstract semantic level, based on which fine-level content features are adaptively learned for more accurate multi-level content correspondences. Extensive experiments demonstrate the superiority of our versatile framework over state-of-the-art methods in robust, high-quality and diversified translations, even for challenging and distant domains. Code is available at* https://github.com/williamyang1991/GP-UNIT.

1. Introduction

Unsupervised image-to-image translation (UNIT) aims to translate images from one domain to another without paired data. Mainstream UNIT methods assume a bijection between domains and exploit cycle-consistency [43] to build cross-domain mappings. Though good results are achieved in simple cases like horse-to-zebra translations, such assumption is often too restrictive for more general heterogeneous and asymmetric domains in the real world. The performance of existing methods often degrades dramatically in translations with large cross-domain shape and appearance discrepancies such as translating human faces to animal faces, limiting their practical applications.

Translating across domains with large discrepancies requires one to establish the translation at a higher semantic level [40]. For instance, to translate a human face to a cat face, one can use the more reliable correspondence of facial components such as the eyes between a human and a cat rather than on the local textures. In the more extreme case of distant domains, such as animals and man-made objects, a translation is still possible if their correspondence can be determined at a higher abstract semantic level, for example through affirming the frontal orientation of an object or the layout of an object within the image.

Establishing translations at different semantic levels demands a UNIT model's ability to find accurate correspondences of different semantic granularity. This requirement is clearly too stringent since training a translation model with such a capability requires complex ground truth correspondences that either do not exist or are infeasible to collect.

Figure 2. Generative space of BigGAN [3]. Objects of different classes generated from the same latent code have a high degree of content correspondences.

In this work, we overcome the aforementioned problems through a novel use of generative prior and achieve promising results as shown in Fig. 1. Specifically, we show that a class-conditional GAN, such as BigGAN [3], provides powerful hints on how different objects are linked – objects of different classes generated from the same latent code have a high degree of content correspondences (Fig. 2). Through generating pairs of such cross-domain images, we can mine the unique prior of the class-conditional GAN and use them to guide an image translation model in building effective and adaptable content mappings across various classes (we will use "domain" instead of "class" hereafter).

However, such prior is not immediately beneficial to UNIT. BigGAN, by nature, covers a large number of domains, which makes it an ideal choice of prior for our problem to achieve translation between multiple domains. However, the coverage of many domains inevitably limits the quality and intra-domain diversity of the captured distribution of each domain. Without a careful treatment, such a limitation will severely affect the performance of UNIT in generating high-quality and diverse results.

To overcome the problem above, we decompose a translation task into coarse-to-fine stages: 1) generative prior distillation to learn robust cross-domain correspondences at a high semantic level and 2) adversarial image translation to build finer adaptable correspondences at multiple semantic levels. In the first stage, we train a content encoder to extract disentangled content representation by distilling the prior from the content-correlated data generated by BigGAN. In the second stage, we apply the pre-trained content encoder to the specific translation task, independent of the generative space of BigGAN, and propose a dynamic skip connection module to learn adaptable correspondences, so as to yield plausible and diverse translation results.

To our knowledge, this is the first work to employ Big-GAN generative prior for unsupervised[1] image-to-image

translation. In particular, we propose a versatile Generative Prior-guided UNsupervised Image-to-image Translation framework (**GP-UNIT**) to expand the application scenarios of previous UNIT methods that mainly handles close domains. Our framework shows positive improvements over previous cycle-consistency-guided frameworks in: 1) capturing coarse-level correspondences across various heterogeneous and asymmetric domains, beyond the ability of cycle-consistency guidance; 2) learning fine-level correspondences applicable to various tasks adaptively; and 3) retaining essential content features in the coarse-to-fine stages, avoiding artifacts from the source domain commonly observed in cycle reconstruction.

In summary, our contributions are threefold:

- We propose a versatile GP-UNIT framework that promotes the overall quality and applicability of UNIT with BigGAN generative prior.
- We present an effective way of learning robust correspondences across non-trivially distant domains at a high semantic level via generative prior distillation.
- We design a novel coarse-to-fine scheme to learn cross-domain correspondences adaptively at different semantic levels.

2. Related Work

Unsupervised image-to-image translation. To learn the mapping between two domains without supervision, CycleGAN [43] proposes a novel cycle consistency constraint to build a bi-directional relationship between domains. To better capture domain-invariant features, representation disentanglement has been investigated extensively in UNIT, where a content encoder and a style encoder [6, 7, 17, 18, 29, 30] are usually employed to extract domain-invariant content features and domain-specific style features, respectively. However, learning a disentangled representation between two domains with drastic differences is non-trivial. To cope with the large visual discrepancy, COCO-FUNIT [37] designs a content-conditioned style encoder to prevent the translation of task-irrelevant appearance information. TGaGa [40] uses landmarks to build geometry mappings. TraVeLGAN [2] proposes a siamese network to seek shared semantic features across domains, and U-GAT-IT [25] leverages an attention module to focus on important regions distinguishing between two domains. These methods struggle to seek powerful and balanced domain-related representation for specific domains so are less adaptive to the various translation tasks, inevitably failing in certain cases. Different from these methods, we propose a new coarse-to-fine scheme – coarse-level cross-domain content correspondences at a highly abstract seman-

---

[1]Following the definition in Liu *et al*. [29], we call our method unsuper-

vised since our method and the pre-trained BigGAN only use the marginal distributions in individual domains without any explicit cross-domain correspondence supervision.
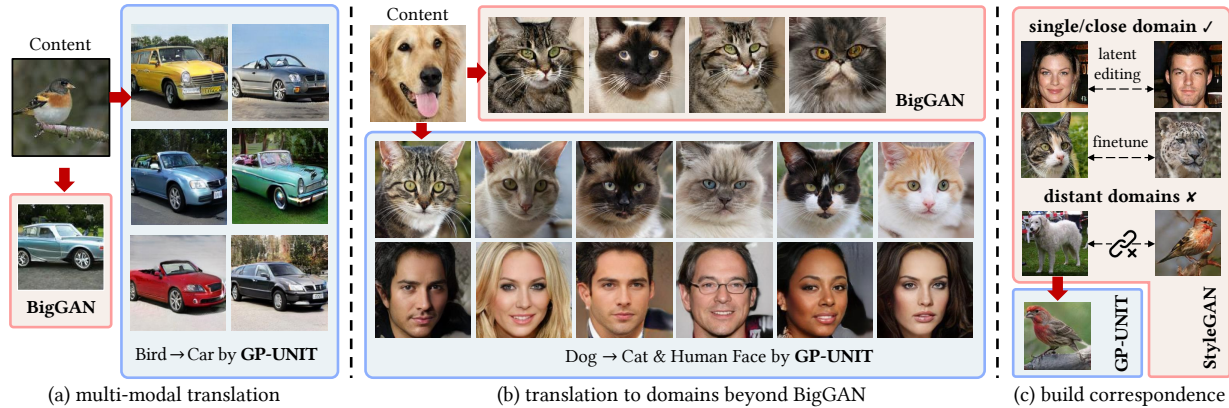
Figure 3. Comparison of the generative spaces of BigGAN, StyleGAN and GP-UNIT. GP-UNIT realizes multi-modal translation, generate cats and human faces beyond ImageNet, and build robust mappings between distant domains. StyleGAN images are from [26, 38]

tic level are first built, based on which fine-level correspondences adaptive to the task are gradually learned. Such a scheme empowers us to build robust mappings to handle various tasks.

Adversarial image generation. Generative Adversarial Network (GAN) [11] introduces a discriminator to compete with the generator to adversarially approximate the real image distribution. Among various models, StyleGAN [23, 24] has shown promising results. Many works [4, 8, 19, 33, 42] exploit the generative prior from StyleGAN to ensure superior image quality by restricting the modulated image to be within the generative space of StyleGAN. However, StyleGAN is an unconditional GAN that is limited to a single domain or close domains [26]. BigGAN [3] is able to synthesize images in different domains but at the expense of quality and intra-domain diversity. Thus it is not straightforward to exploit BigGAN prior following the aforementioned works. To circumvent this limitation, in this paper, we distill the generative prior from content-correlated data generated by BigGAN and apply it to the image translation task to generate high-quality images.

3. Generative Prior Distillation

3.1. Cross-Domain Correspondences Prior

Our framework is motivated by the following observation [1, 13] – objects generated by BigGAN, despite originating from different domains, share high content correspondences when generated from the same latent code. Figure 2 shows the generative space of BigGAN characterized by three latent codes (z_1, z_2, z_3) across five domains. For each latent code, fine-grained correspondences can be observed between semantically related dogs and cats, such as the face features and body postures. For birds and vehicles, which are rather different, one can also observe coarse-level correspondences in terms of orientation and layout.

The interesting phenomenon suggests that *there is an*

inherent content correspondence at a highly abstract semantic level regardless of the domain discrepancy in the BigGAN generative space. In particular, objects with the same latent code share either the same or very similar abstract representations in the first few layers, based on which domain-specific details are gradually added.

In this paper, we exploit this generative prior for building robust mappings and choose BigGAN for its rich cross-domain prior. Nevertheless, its generative space is limited in **quality** and **diversity** for our purpose. In terms of quality, BigGAN sometimes generates unrealistic objects, such as the dog body with z_2 in Fig. 2. As for diversity, first, the space lacks intra-domain variation, *e.g.*, the diversity in textures of a dog or colors of a bird in the same domain are pretty limited. Using such prior in UNIT will overfit the model to limited appearances. Second, the BigGAN generative space is limited to 1,000 domains of ImageNet [36], which is inadequate for an actual UNIT. For instance, it only has four kinds of domestic cats as in Fig. 3(b) and excludes the important domain of human faces.

StyleGAN is not suitable for our task despite its generative space is of high quality and diversity. This is because it is limited to a single domain, and thus it is mainly applied to attribute transfer within one domain via latent editing [8, 33, 42]. Recently, cross-domain translations on StyleGAN have been achieved via finetuning [26, 34], but this still assumes a small distance between the models of the source and target domains and is therefore still restricted to close domains. The assumption makes StyleGAN prior less applicable to more complex translation tasks.

Our framework solves the above problems by distilling a general generative prior from BigGAN instead of directly constraining the latent or image space. It enables us to design and train the translation module independently. Therefore, we can realize multi-modal translation (Fig. 3(a)), generalize to classes beyond ImageNet (Fig. 3(b)) and build robust mappings between distant domains (Fig. 3(c)). Next,

(a) Stage I: Generative Prior Distillation   **(b) Stage II: Adversarial Unsupervised Image Translation**
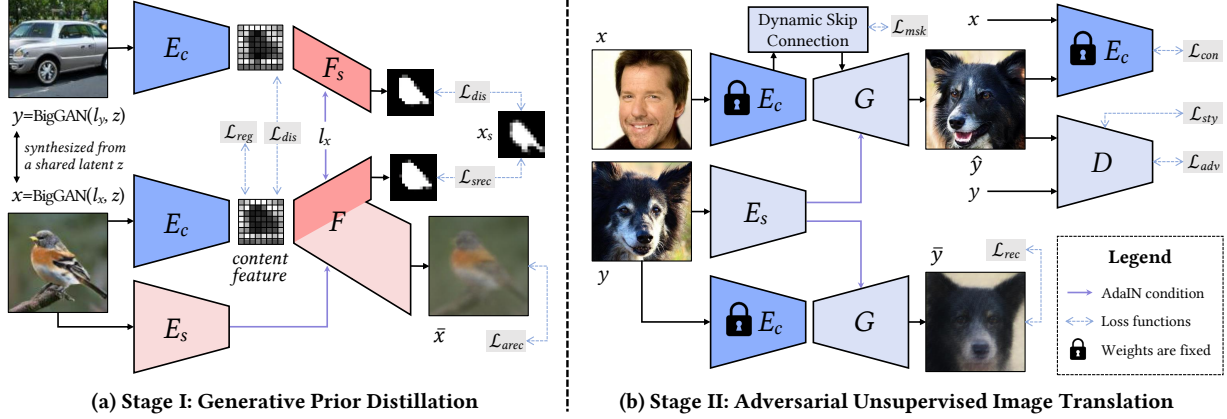
Figure 4. Overview of the proposed GP-UNIT. In the first stage, we use a content encoder E_c to extract shared coarse-level content features between a pair correlated images (x, y) generated by BigGAN from a common random latent code in two random domains. In the second stage, we build our translation network based on the content encoder E_c in the first stage. For simplicity, we omit the classifier C.

we detail how we distill the prior.

3.2. Prior Distillation with Content Encoder

Given the correlated images (x, y) generated by Big-GAN from a common random latent code in two random domains $\mathcal{X}$ and $\mathcal{Y}$, our main goal is to train a content encoder E_c to extract their shared coarse-level content features, which can be used to reconstruct their shape and appearance. Figure 4(a) illustrates this autoencoder pipeline for generative prior distillation.

Specifically, we use a decoder F to recover the appearance x based on its content feature $E_c(x)$, style feature $E_s(x)$ extracted by a style encoder E_s and the domain label l_x. We further exploit the shallow layers F_s of F to predict the shape of x (*i.e.*, instance segmentation map x_s, which is extracted from x by HTC [5]) based on $E_c(x)$ and l_x. We find such auxiliary prediction eases the training on hundreds of domains. Besides the shape and appearance reconstruction, we further regularize the content feature in three ways for disentanglement: 1) x and y should share the same content feature; 2) We introduce a classifier C with a gradient reversal layer R [10] to make the content feature domain-agnostic; 3) We limit $E_c(x)$ to one channel to eliminate domain information [39] and add Gaussian noise of a fixed variance for robustness. Our objective function is:

$$\min_{E_c, E_s, F, C} \mathcal{L}_{arec} + \mathcal{L}_{srec} + \mathcal{L}_{dis} + \mathcal{L}_{reg}, \quad (1)$$

where $\mathcal{L}_{arec}$ is the appearance reconstruction loss measuring the L_2 and perceptual loss [20] between $\bar{x} = F(E_c(x), E_s(x), l_x)$ and x. The shape reconstruction loss, $\mathcal{L}_{srec}$, is defined as

$$\mathcal{L}_{srec} = \lambda_s \mathbb{E}_x[\|F_s(E_c(x), l_x) - x_s\|_1]. \quad (2)$$

The binary loss $\mathcal{L}_{dis}$ with paired inputs narrows the distance between the content features of x and y. In addition, we

would like to recover the shape of x with the content feature of y, which simulates translations:

$$\mathcal{L}_{dis} = \mathbb{E}_{(x,y)}[\|E_c(x) - E_c(y)\|_1 + \lambda_s\|F_s(E_c(y), l_x) - x_s\|_1].$$

Finally, $\mathcal{L}_{reg}$ guides C to maximize the classification accuracy and pushes E_c to confuse C, so that the content feature is domain-agnostic. An L_2 norm is further applied to the content feature:

$$\mathcal{L}_{reg} = \mathbb{E}_x[-l_x \log C(R(E_c(x)))] + \lambda_r \mathbb{E}_x[\|E_c(x)\|_2]. $$

For unary losses of $\mathcal{L}_{arec}$, $\mathcal{L}_{srec}$ and $\mathcal{L}_{reg}$, we also use real images of ImageNet [36] and CelebA-HQ [21] for training to make E_c more generalizable.

4. Adversarial Image Translation

Given a fixed content encoder E_c pre-trained in the first stage, we build our translation network following a standard style transfer paradigm in the second stage. Thanks to the pre-trained E_c that provides a good measurement for content similarity, our framework does not need cycle training.

As shown in Fig. 4(b), our translation network receives a content input $x \in \mathcal{X}$ and a style input $y \in \mathcal{Y}$. The network extracts their content feature $E_c(x)$ and style feature $E_s(y)$, respectively. Then, a generator G modulates $E_c(x)$ to match the style of y via AdaIN [16], and finally produces the translated result $\hat{y} = G(E_c(x), E_s(y))$. The realism of $\hat{y}$ is reinforced through an adversarial training with a discriminator D,

$$\mathcal{L}_{adv} = \mathbb{E}_y[\log D(y)] + \mathbb{E}_{x,y}[\log(1 - D(\hat{y}))]. \quad (3)$$

In addition, $\hat{y}$ is required to fit the style of y, while preserving the original content feature of x, which can be formulated as a style loss $\mathcal{L}_{sty}$ and a content loss $\mathcal{L}_{con}$,

$$\mathcal{L}_{sty} = \mathbb{E}_{x,y}[\|f_D(\hat{y}) - f_D(y)\|_1], \quad (4)$$

$$\mathcal{L}_{con} = \mathbb{E}_{x,y}[\|E_c(\hat{y}) - E_c(x)\|_1], \quad (5)$$

where f_D is the style feature defined as the channel-wise mean of the middle layer feature of D following the style definition in [16].

4.1. Dynamic Skip Connection

Domains that are close semantically would usually exhibit fine-level content correspondences that cannot be characterized solely by the abstract content feature. To solve this problem, we propose a dynamic skip connection module, which passes middle layer features f_E from E_c to G and predicts masks m to select the valid elements for establishing fine-level content correspondences.

Our dynamic skip connection is inspired by the GRU-like selective transfer unit [28]. Let the superscript l denote the layer of G. The mask m^l at layer l is determined by the encoder feature f_E^l passed to the same layer and a hidden state h^{l-1} at the last layer. Specifically, we first set the first hidden state $h^0 = E_c(x)$ and use the upsampling convolution to match the dimension of h^{l-1} with f_E^l as $\hat{h}^{l-1} = \sigma(W_h^l \circ \uparrow h^{l-1})$, where $\uparrow$, $\circ$ and W_h^l are the upsample operator, convolution operator and convolution weights, respectively. The activation layer is denoted as σ. Then, our module at layer l updates the hidden state h^l and the encoder feature $\hat{f}_E^l$, and fuses $\hat{f}_E^l$ with the generator feature f_G^l with the predicted mask m^l:

$$r^l = \sigma(W_r^l \circ [\hat{h}^{l-1}, f_E^l]), \quad m^l = \sigma(W_m^l \circ [\hat{h}^{l-1}, f_E^l]),$$
$$h^l = r^l \hat{h}^{l-1}, \quad \hat{f}_E^l = \sigma(W_E^l \circ [h^l, f_E^l]),$$
$$f^l = (1-m^l)f_G^l + m^l \hat{f}_E^l,$$

where $[\cdot, \cdot]$ denotes concatenation. m^l has the same dimension of f_G^l, serving both channel attention and spatial attention. Moreover, we apply L_1 norm to m^l to make it sparse,

$$\mathcal{L}_{msk} = \sum_l \mathbb{E}_x[\|m^l\|_1], \tag{6}$$

so that only the most useful content cues from the source domain are selected.

Full objectives. Combining the aforementioned losses, our full objectives take the following form:

$$\min_{G,E_s} \max_D \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{sty} + \lambda_3 \mathcal{L}_{msk} + \lambda_4 \mathcal{L}_{rec}.$$

A reconstruction loss $\mathcal{L}_{rec}$ is added to measure the L_1 and perceptual loss [20] between y and $\bar{y} = G(E_c(y), E_s(y))$. Intuitively, we would like the learned style feature of an image to precisely reconstruct itself with the help from its content feature, which stabilizes the network training.

Style sampling. To sample latent style features directly for multi-modal generation without the style images, we follow the post-processing of [32] to train a mapping network to map the unit Gaussian noise to the latent style distribution using a maximum likelihood criterion [15]. Please refer to [15] for the details.

5. Experimental Results

Dataset. In the first stage, we prepare both synthesized data and real data. For synthesized data, we use the official BigGAN [3] to generate correlated images associated by random latent codes for each of the 291 domains including animals and vehicles. After filtering low-quality ones, we finally obtain 655 images per domain that are linked across all domains, 600 of which are for training. We denote this dataset as *synImageNet-291*. For real data, we apply HTC [5] to ImageNet [36] to detect and crop the object regions. Each domain uses 600 images for training. We denote this dataset as *ImageNet-291*. Besides, 29K face images of CelebA-HQ [21,31] are also included for training.

In the second stage, we perform evaluation on four translation tasks. 1) Male ↔ Female on 28K training images of CelebA-HQ [21,31]. 2) Dog ↔ Cat on AFHQ [7], with 4K training images per domain. 3) Human Face ↔ Cat on 4K AFHQ images and 29K CelebA-HQ images. 4) Bird ↔ Dog or Car: Four classes of birds, four classes of dogs and four classes of cars in *ImageNet-291* are used. Every four classes form a domain with 2.4K training images. Here, we use Bird↔ Car as the extreme case to test to what extent GP-UNIT can handle for stress testing.

Network training. We set $\lambda_s = 5$, $\lambda_r = 0.001$, $\lambda_1 = \lambda_3 = \lambda_4 = 1$ and $\lambda_2 = 50$. For Cat → Human Face, we use an additional identity loss [9] with weight 1 to preserve the identity of the reference face following [35]. Dynamic skip connections are applied to the 2nd layer ($l = 1$) and the 3rd layer ($l = 2$) of G. Except for Male ↔ Female, we do not use dynamic skip connections to compute $\mathcal{L}_{rec}$ (by setting m^l to an all-zero tensor), which are discussed in Sec. 5.3.

5.1. Comparison with the State of the Arts

Qualitative comparison. We perform visual comparison to six state-of-the-art methods in Fig. 5 and Fig. 6. As shown in Fig. 5, cycle-consistency-guided U-GAT-IT [25], MU-NIT [17] and StarGAN2 [7] rely on the low-level cues of the input image for bi-directional reconstruction, which leads to some undesired artifacts, such as the distorted cat face region that corresponds to the dog ears, and the ghosting dog legs in the generated bird images. Meanwhile, TraVeL-GAN [2] and COCO-FUNIT [37] fail to build proper content correspondences for Human Face ↔ Cat and Bird ↔ Car. By comparison, our method is comparable to the above methods on Male ↔ Female task and show consistent superiority on other challenging tasks. In Fig. 6, we compare our model to TGaGa [40], which also deals with large geometric deformations on exemplar-guided translation. TGaGa produces blurry results and fails to match the example appearance precisely, *e.g.*, all the generated faces look alike, except for the color changes. GP-UNIT surpasses TGaGa in both vivid details and style consistency.
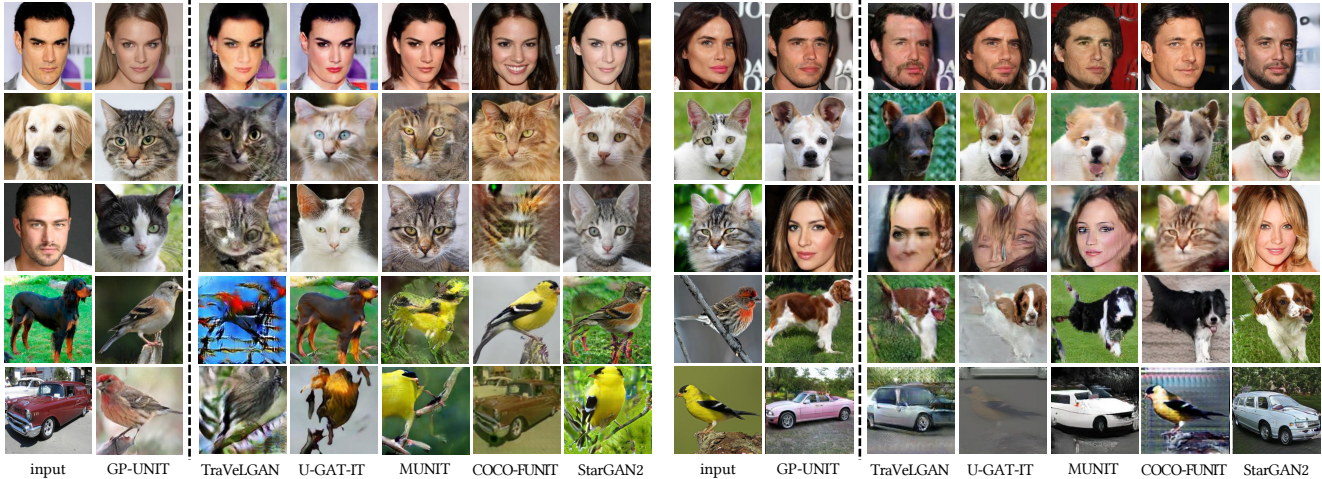
Figure 5. Visual comparison with TraVeLGAN [2], U-GAT-IT [25], MUNIT [17], COCO-FUNIT [37] and StarGAN2 [7]. GP-UNIT consistently outperforms on all tasks and demonstrates greater superiority as the task becomes more challenging (from top to bottom).

Table 1. Quantitative comparison. We use FID and Diversity with LPIPS to evaluate the quality and diversity of the generated images.

Task	Male ↔ Female		Dog ↔ Cat		Human Face ↔ Cat		Bird ↔ Dog		Bird ↔ Car		Average	
Metric	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity	FID	Diversity
TraVeLGAN	66.60	–	58.91	–	85.28	–	169.98	–	164.28	–	109.01	–
U-GAT-IT	29.47	–	38.31	–	110.57	–	178.23	–	194.05	–	110.12	–
MUNIT	22.64	0.37	80.93	0.47	56.89	**0.53**	217.68	0.57	121.02	0.60	99.83	0.51
COCO-FUNIT	39.19	0.35	97.08	0.08	236.90	0.33	30.27	0.51	207.92	0.12	122.27	0.28
StarGAN2	**14.61**	**0.45**	22.08	0.45	**11.35**	0.51	20.54	0.52	29.28	0.58	19.57	0.50
GP-UNIT	14.63	0.37	**15.29**	**0.51**	13.04	0.49	**11.29**	**0.60**	**13.93**	**0.61**	**13.64**	**0.52**

Figure 6. Comparing exemplar-guided translation with TGaGa[2]. GP-UNIT surpasses TGaGa in vivid details and style consistency.

Quantitative comparison. We follow [7, 40] to perform quantitative comparison in terms of quality and diversity. FID [14] and LPIPS [41] are used to evaluate the image quality of generated results against the real data and the output diversity, respectively. For methods supporting multi-modal transfer (MUNIT, COCO-FUNIT, StarGAN2, GP-UNIT), we generate 10 paired translation results per test

Table 2. User preference scores in terms of content consistency and overall preference. Best scores are marked in bold.

Metric	Content Consistency	Overall Preference
TraVeLGAN	0.012	0.006
U-GAT-IT	0.076	0.050
MUNIT	0.045	0.033
COCO-FUNIT	0.065	0.044
StarGAN2	0.199	0.171
GP-UNIT	**0.603**	**0.696**

image from randomly sampled latent codes or exemplar images to compute their average LPIPS distance. The quantitative results averaged over all test images are reported in Table 1, which are consistent with Fig. 5, *i.e.*, our method is comparable or superior to the compared methods, and the advantage becomes more distinct on difficult tasks, obtaining the best overall FID and diversity. We find GP-UNIT tends to preserve the background of the input image. This property does not favor diversity, but might be useful in some applications. Although StarGAN2 yields realistic human faces (best FID) on Cat → Human Face, it ignores the pose correspondences with the input cat faces (lower content consistency than GP-UNIT), as in Fig. 5.

We further conduct a user study to evaluate the input-output content consistency and overall translation performance. A total of 25 subjects participate in this study to select what they consider to be the best results from the six

---

[2]At the time of this submission, the code and training data of TGaGa are not released. We directly use the test and result images kindly provided by the authors of TGaGa. Since the training data of GP-UNIT and TGaGa does not match, this comparison is for visual reference only.

| content | style | w/o prior | full model | | content | style | w/o DSC | full model | m^1 #135 | m^1 #301 |

(a) effect of the generative prior   (b) effect of the dynamic skip connection

Figure 7. Ablation study on the generative prior and the dynamic skip connection.

| input | m^1=**0**, m^2=**0** | m^2=**0** | full model | | input | m^1=**0**, m^2=**0** | m^2=**0** | full model |

| enlarged regions | $E_c(x)$ | m^1 #305 | m^2 #169 | | $E_c(x)$ | m^1 #85 | m^2 |

(a) multi-level correspondences between cat and tiger   (b) multi-level correspondences between human and dog
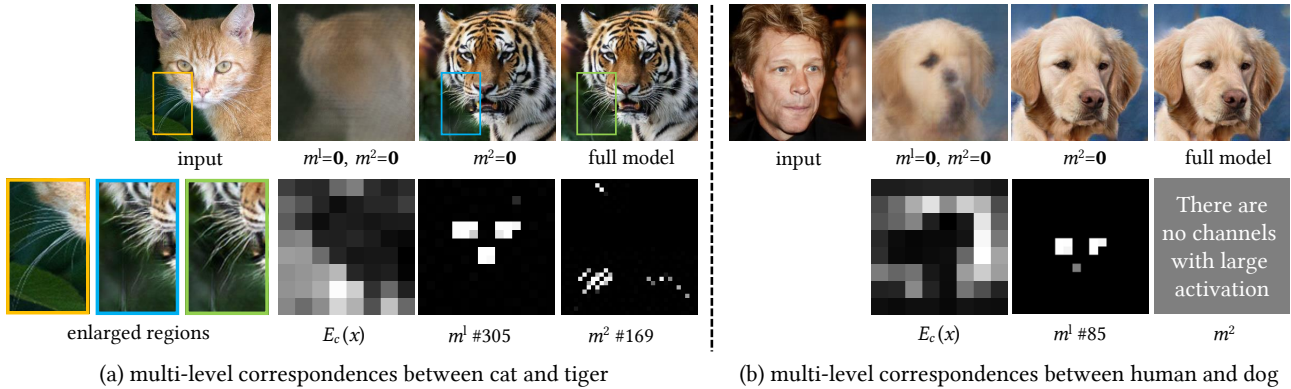
Figure 8. Our framework learns multi-level content correspondences that are robust and adaptable to different translation tasks.

methods, and a total of 2,500 selections on 50 groups of results are tallied. Table 2 summarizes the average preference scores, where the proposed method receives notable preference for both content consistency and overall performance.

5.2. Ablation Study

Generative prior distillation. As shown in Fig. 7(a), if we train our content encoder from scratch along with all other subnetworks in the second stage, like most image translation frameworks, this variant fails to preserve the content features such as the eye position. By comparison, our pre-trained content encoder successfully exploits the generative prior to build effective content mappings. It also suggests the necessity of the coarse-level content feature, only based on which valid finer-level features can be learned. Hence, the generative prior is the key to the success of our coarse-to-fine scheme of content correspondence learning.

Dynamic skip connection. As shown in Fig. 7(b), without dynamic skip connections (DSC), the model cannot keep the relative position of the nose and eyes as in the content images. We show that the 135th and 301st channels of the mask m^1 predicted by our full model effectively locate these features for accurate content correspondences.

Multi-level cross-domain correspondences. Figure 8 analyzes the learned multi-level correspondences. The most abstract $E_c(x)$ only gives layout cues. If we solely use $E_c(x)$ (by setting both masks m^1 and m^2 to all-zero tensors), the resulting tiger and dog faces have no details. Meanwhile, m^1 focuses on mid-level details like the nose and eyes of cat face in the 305th channels, and eyes of human face in the 85the channels, which is enough to generate a realis-

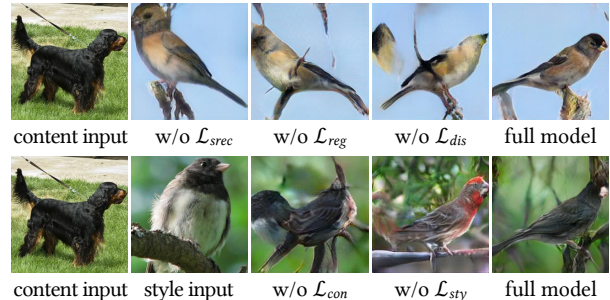| content input | w/o $\mathcal{L}_{srec}$ | w/o $\mathcal{L}_{reg}$ | w/o $\mathcal{L}_{dis}$ | full model |
| content input | style input | w/o $\mathcal{L}_{con}$ | w/o $\mathcal{L}_{sty}$ | full model |

Figure 9. Effects of the loss terms.

tic result with $E_c(x)$. Finally, m^2 pays attention to subtle details like the cat whiskers in the 169th channel for close domains. Therefore, our full multi-level content features enable us to simulate the extremely fine-level long whiskers in the input. As expected, such kind of fine-level correspondences are not found between more distant human and dog faces, preventing the unwanted appearance influence from the source domain (*e.g.*, clothes in the generated cat faces in Fig. 5). Note that such reasonable and adaptable semantic attentions are learned merely via the generation prior, without any explicit correspondence supervision.

Loss functions. Figure 9 studies the effects of the loss terms. In Stage I, $\mathcal{L}_{srec}$ is the key to learn correct content features, or correspondence is not built. $\mathcal{L}_{reg}$ makes content features more sparse to improve robustness to unimportant domain-specific details. $\mathcal{L}_{dis}$ finds domain-shared features to prevent the output from affected by objects from the source domain like the dog tail. In Stage II, $\mathcal{L}_{con}$ helps strengthen the pose correspondence while $\mathcal{L}_{sty}$ makes the output better match the style of the exemplar image.
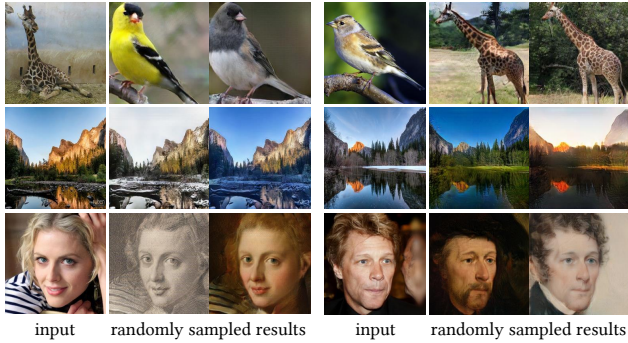
input | randomly sampled results | input | randomly sampled results

Figure 10. Applicability to domains beyond BigGAN: (*top*) Giraffe ↔ Bird, (*middle*) Summer ↔ Winter, (*bottom*) Face → Art.
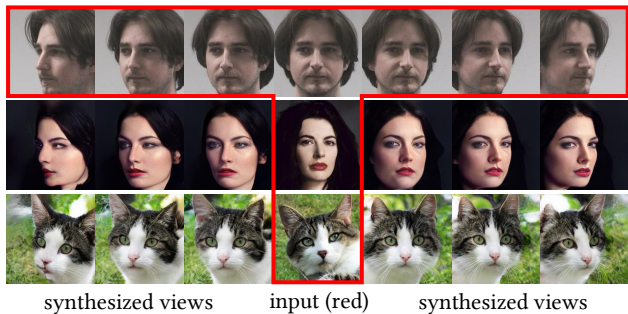

synthesized views | input (red) | synthesized views

Figure 11. Unseen view synthesis.

5.3. More Results

Generalization to domains beyond BigGAN. Figure 10 shows three applications of species translation, season transfer and facial stylization. Even if MS-COCO giraffes [27], Yosemite landscapes [17] and Art portraits [22] are not within the ImageNet 1,000 classes and are not observed by the content encoder in Stage I, our method can well support these domains and generate realistic results.

Unseen view synthesis. Our exemplar-guided framework allows unseen view synthesis. Figure 11 shows our synthesized realistic human and cat faces in various pan angles according to the reference faces from the Head Pose Image Database [12]. To transfer identity and prevent low-level content correspondence, we add an identity loss [9] and do not use DSC for $\mathcal{L}_{rec}$. We further show the usage of DSC for $\mathcal{L}_{rec}$ can flexibly control the facial attribute to be transferred in Fig. 12. Using DSC for $\mathcal{L}_{rec}$ helps preserve the identity of the content face, which is suitable for gender and color transfer. Meanwhile, using identity loss without DSC for $\mathcal{L}_{rec}$, most attributes of the style face except pose can be transferred, which is suitable for pose transfer.

5.4. Limitations

Figure 13 gives three typical failure cases of our method. First, it is hard to learn certain semantic correspondences solely from the object appearance, such as which side of a car is its front. We observe that a bird tail is often trans-
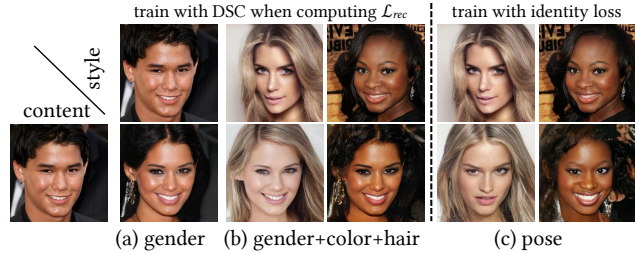

(a) gender (b) gender+color+hair (c) pose

Figure 12. Flexible multi-level attribute transfer.

input | result | input | result | content | style | result

Figure 13. Failure cases caused by (*left*) the lack of semantical supervision, (*middle*) imbalanced training data, and (*right*) scale mismatch of the content and style objects.

lated into the front of a car since they are both the thinner part of the objects. Second, our method fails to generate a bird sharing the same head direction as the dog, due to the lack of training images of birds looking directly at the camera. Therefore, special attention should be taken when applying this method to applications where the possible data imbalance issue might lead to biased results towards minority groups in the dataset. Finally, when the objects in the content and style images have very different scales, some appearance features cannot be rendered correctly.

6. Conclusion and Discussion

In this paper, we explore the use of GAN generative prior to build a versatile UNIT framework. We show that the proposed two-stage framework is able to characterize content correspondences at a high semantic level for challenging multi-modal translations between distant domains. An advantage is that such content correspondences can be discovered with only domain supervision (*i.e.*, only knowing the domain each image belongs to). We further find in Sec. 5.2 that fine-level correspondences are learned merely via a generation task. This might suggest an intriguing behavior of deep neural networks to automatically find and integrate shared appearance features across domains in a coarse-to-fine manner in order to reconstruct various objects. It poses a potential of Learning by Generation: building object relationships by generating and transforming them. Another interesting topic is to learn semantic correspondences beyond object appearance, such as the frontal side of an object discussed in Sec. 5.4. One possible direction is semi-supervised learning where a small amount of data could be labeled to specify the semantic correspondences.

References

[1] Zaid Alyafeai. Bigganex: A dive into the latent space of biggan. *The Gradient*, 2018. https://thegradient.pub/bigganex-a-dive-into-the-latent-space-of-biggan/. 3

[2] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8983–8992, 2019. 2, 5, 6

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *Proc. Int'l Conf. Learning Representations*, 2019. 2, 3, 5

[4] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021. 3

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 4, 5

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018. 2

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2, 5, 6

[8] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 3

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5, 8

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. IEEE Int'l Conf. Machine Learning*, pages 1180–1189. PMLR, 2015. 4

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 3

[12] Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial features. In *ICPR International Workshop on Visual Observation of Deictic Gestures*. Citeseer, 2004. 8

[13] Erik Härkönen, Aaron Hertzman, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems*, 2020. 3

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6629–6640, 2017. 6

[15] Yedid Hoshen, Ke Li, and Jitendra Malik. Non-adversarial image synthesis with generative latent nearest neighbors. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 5811–5819, 2019. 5

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. Int'l Conf. Computer Vision*, pages 1510–1519, 2017. 4, 5

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. European Conf. Computer Vision*, pages 172–189. Springer, 2018. 2, 5, 6, 8

[18] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *Proc. European Conf. Computer Vision*, pages 206–222. Springer, 2020. 2

[19] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proc. Int'l Conf. Computer Vision*, pages 13799–13808, 2021. 3

[20] Justin Johnson, Alexandre Alahi, and Fei Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conf. Computer Vision*, pages 694–711. Springer, 2016. 4, 5

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *Proc. Int'l Conf. Learning Representations*, 2018. 4, 5

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems*, 2020. 8

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3

[25] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *Proc. Int'l Conf. Learning Representations*, 2019. 2, 5, 6

[26] Sam Kwong, Jialu Huang, and Jing Liao. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia*, 2021. 3

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conf. Computer Vision*, pages 740–755. Springer, 2014. 8

[28] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 3673–3682, 2019. 5

[29] Ming Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 2

[30] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proc. Int'l Conf. Computer Vision*, pages 10551–10560, 2019. 2

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. Int'l Conf. Computer Vision*, pages 3730–3738, 2015. 5

[32] Moustafa Meshry, Yixuan Ren, Larry S Davis, and Abhinav Shrivastava. Step: Style-based encoder pre-training for multi-modal image synthesis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021. 5

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. 2021. arXiv:2103.17249. 3

[34] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020. 3

[35] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021. 5

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int'l Journal of Computer Vision*, 115(3):211–252, 2015. 3, 4, 5

[37] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In *Proc. European Conf. Computer Vision*. Springer, 2020. 2, 5, 6

[38] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 3

[39] Vadim Sushko, Juergen Gall, and Anna Khoreva. One-shot gan: Learning to generate samples from single images and videos. 2021. arXiv:2103.13389. 4

[40] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8012–8021, 2019. 1, 2, 5, 6

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[42] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proc. European Conf. Computer Vision*, pages 592–608. Springer, 2020. 3

[43] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. Int'l Conf. Computer Vision*, pages 2242–2251, 2017. 1, 2