One-for-All: Towards Universal Domain Translation with a Single StyleGAN

Yong Du, *Member, IEEE*, Jiahui Zhan, Xinzhe Li, Junyu Dong, *Member, IEEE*, Sheng Chen, *Life Fellow, IEEE*, Ming-Hsuan Yang, *Fellow, IEEE*, and Shengfeng He, *Senior Member, IEEE*

Abstract—In this paper, we propose a novel translation model, UniTranslator, for transforming representations between visually distinct domains under conditions of limited training data and significant visual differences. The main idea behind our approach is leveraging the domain-neutral capabilities of CLIP as a bridging mechanism, while utilizing a separate module to extract abstract, domain-agnostic semantics from the embeddings of both the source and target realms. Fusing these abstract semantics with target-specific semantics results in a transformed embedding within the CLIP space. To bridge the gap between the disparate worlds of CLIP and StyleGAN, we introduce a new non-linear mapper, the CLIP2P mapper. Utilizing CLIP embeddings, this module is tailored to approximate the latent distribution in the StyleGAN's latent space, effectively acting as a connector between these two spaces. The proposed UniTranslator is versatile and capable of performing various tasks, including style mixing, stylization, and translations, even in visually challenging scenarios across different visual domains. Notably, UniTranslator generates high-quality translations that showcase domain relevance, diversity, and improved image quality. UniTranslator surpasses the performance of existing general-purpose models and performs well against specialized models in representative tasks. The source code is available at https://zhanjiahui.github.io/UniTranslator/.

Index Terms—Generative Adversarial Networks, Image-to-Image Translation, GAN Embedding.

1 Introduction

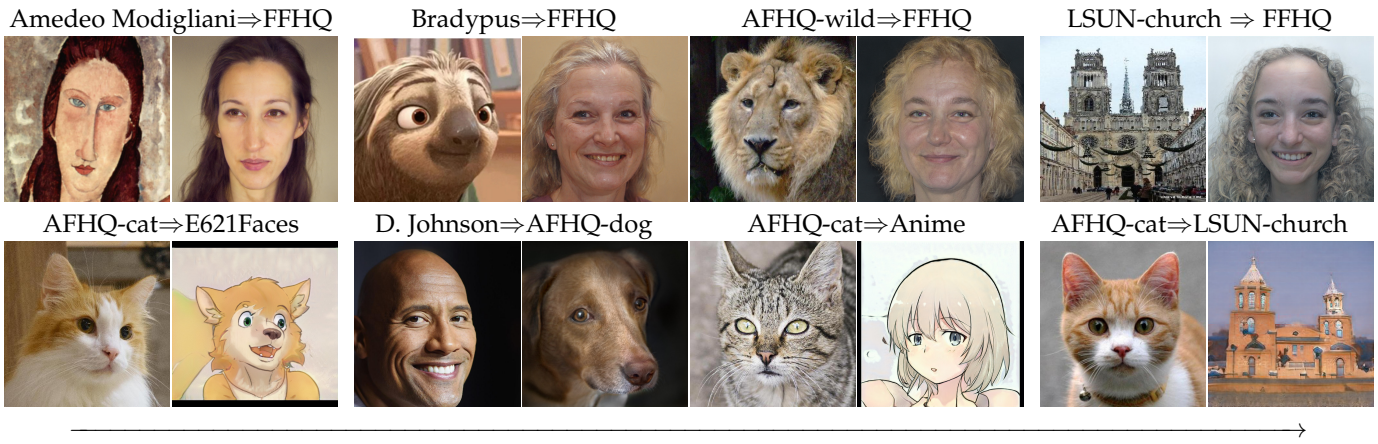
Recent generative models are developed with a growing emphasis on universality, aiming to enhance the real-world applicability in solving complex challenges [1], [2], [3]. Significant advances have been made in visual domain translation [4], [5], [6], [7], [8], which harnesses the transformation of images by exploiting inherent content correlations across disparate realms. The defining feature of universality for domain translators is their ability to seamlessly convert images from any real-world source domain to a chosen target domain. This pursuit of universality in visual domain translation erodes the barriers segregating different domains and provides invaluable technological support for a wide range of applications. These span from artistic creations such as anthropomorphic or skeuomorphic designs to the entertainment industry, including customized effect generation on various platforms.

Despite significant progress, existing translators encounter several challenges to achieve universality: First, most existing techniques necessitate identifying the source and target domains to learn the mappings between them. However, transitioning between domains frequently entails a laborious model retraining process, curtailing their flexibility and applicability. Although recent models such as StarGAN [9] and StarGAN2 [10] utilize a single model to master many-to-many mappings across all the domains involved, the limitations become apparent in their inability to handle a vast number of domain mappings, thus falling short of achieving true universality.

Second, the training or retraining of existing models typically entails a sizable amount of data. While unsupervised image-to-image translation tasks [11], [12], [13], [14], [15] alleviate the need for paired images, real-world scenarios can still pose formidable challenges, particularly for asymmetric domains. For instance, when leveraging artworks from a specific artist as the source or target domain, the available data may be scant, making it arduous to train the models effectively.

Third, few-shot or diffusion-based domain adaptation approaches [16], [17], [18], [19] have been employed for domain translation, building upon inter-domain correlations. These approaches offer the benefits of requiring minimal data for fine-tuning. However, they are primarily suitable for translating between closely related domains like human→sketch. As the scale of the domain gap increases, their effectiveness in accomplishing robust transformations wanes (see Fig. 2 (c)-(d)). In real-world translation applications, constraining the input domain to be closely aligned with the target domain would be counterproductive. For example, a platform that generates Internet memes may

- *This work is supported by the National Natural Science Foundation of China (No. 62102381, 41927805); Shandong Natural Science Foundation (No. ZR2021QF035); the National Key R&D Program of China (No. 2022ZD0117201); the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097); and National Research Foundation Singapore under the AI Singapore Programme (AISG3-GV-2023-011). The first two authors contribute equally. Corresponding author: Yong Du.*
- *Yong Du, Xinzhe Li, and Junyu Dong are with the School of Computer Science and Technology, Ocean University of China, Qingdao, China. E-mail: csyongdu@ouc.edu.cn, lixinzhe@stu.ouc.edu.cn, dongjunyu@ouc.edu.cn.*
- *Jiahui Zhan is with the School of Computer Science and Technology, Ocean University of China, and Shanghai Jiao Tong University. E-mail: zhanjiahui@stu.ouc.edu.cn.*
- *Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK. Email: sqc@ecs.soton.ac.uk.*
- *Ming-Hsuan Yang is with the University of California at Merced, Yonsei University, and Google. E-mail: mhyang@ucmerced.edu.*
- *Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore. Email: shengfenghe@smu.edu.sg.*



(a) Adjacent domains **(b) Far-off domains** **(c) Intensely far-off domains**

Fig. 1. We introduce UniTranslator, an innovative universal framework for translating across diverse visual domains. It can receive input from any real-world source domain and convert it into a specified target domain, all while ensuring high image quality, domain correspondence, and variability.

need to fulfill users' requests to visualize how a cartoon animal might look as a human or to satisfy their curiosity about what a person with lion-like features would resemble, as depicted in Fig. 1 (b). These scenarios involve translating between domains that are far-off but still related, as both belong to the broader category of living beings. In contrast, transforming oil paintings of people into realistic images (Fig. 1 (a)) involves adjacent domains within the same super-category of "human".

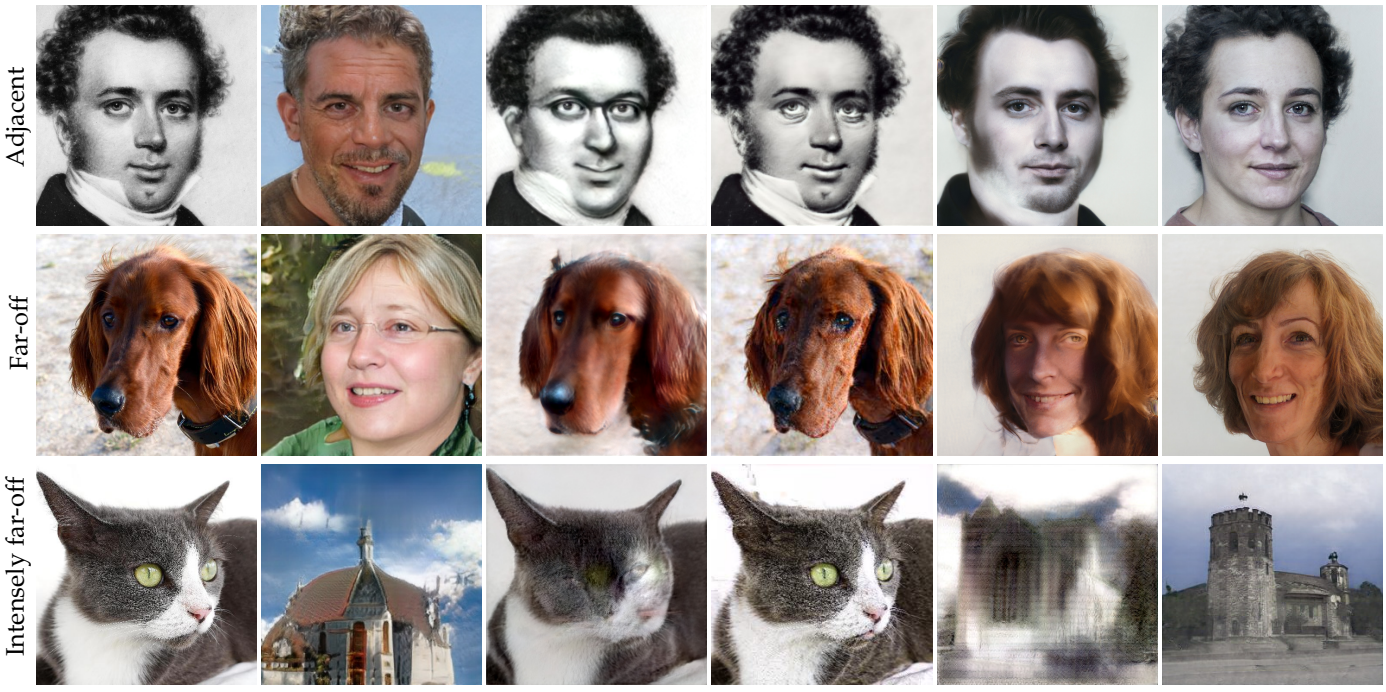
Moreover, creative gaming applications often demand technology capable of handling even more distant domain translations. For instance, in construction games, users might wish to design buildings that resemble their pets, as shown in Fig. 1 (c). This requires translating between living beings and inanimate objects or artificial structures, representing an intensely far-off domain gap. In such cases, image translation techniques that can manage these extreme domain differences are crucial. Therefore, the primary hurdle to achieving universality is the degree of heterogeneity between the source and target domains. Modeling cross-domain correspondences in terms of shape, appearance, and so on poses challenges, especially when there is a significant chasm between domains.

GP-UNIT [20] can handle domain translation with a large gap among existing translators. It uses generative priors distilled from BigGAN [21] to capture coarse-level content correspondences, enabling conversions across highly heterogeneous domains. Nevertheless, GP-UNIT inherits the limitations of BigGAN's latent space, which is not efficiently decoupled, leading to the generation of unrealistic objects. Moreover, due to the intricate division of domains in BigGAN's latent space, the mid-level and fine-level correspondences between various categories of images generated using identical latent codes often fall short. This restriction can affect multi-level cross-domain correspondences in the transformed outcomes with respect to the input (see Fig. 2 (b)). As such, it is essential to develop methods to handle domain translation tasks more efficiently, preserving more natural correspondences and ultimately pushing the boundaries of universality in visual domain translation.

In this paper, we propose the UniTranslator, which allows translating a single input image to a target image in a picked domain. It can generate high-quality, visually continuous, diverse translation results even in scenarios with significant visual differences. Our work bears some resemblance to the optimization-based super-resolution method, PULSE [22], where an input image serves as a reference to guide the search process in the latent space, and the optimized latent code can then be fed into StyleGAN to generate the proper image. The transformation from input to output in PULSE can be regarded as one type of domain translation, while the generative capability of StyleGAN ensures the quality of the translated result. However, PULSE does not perform well in cross-domain translation due to its lack of efficient mechanisms for handling the inherent correlations between heterogeneous domains. As a consequence, the generated results often exhibit a blending of patterns from both domains (see Fig. 2 (e)).

To tackle this issue, we propose a decoupling module in UniTranslator. By integrating descriptive prompts linked to the source image and the intended target domain, this module leverages the language-image alignment proficiency of Contrastive-Language-Image-Pretraining (CLIP) [23] to obtain abstract domain-agnostic semantics. These semantics are then amalgamated with target-specific semantics, refining a target domain embedding that retains cross-domain correlations. However, due to the disparity between the CLIP space and the StyleGAN space, the CLIP embedding may lie beyond the latent domain of StyleGAN, potentially leading to suboptimal conversion results. To overcome this, we devise a mapper that bridges the latent distributions in CLIP and StyleGAN's latent spaces, guided by their statistical properties. This approach effectively translates the meticulously acquired CLIP embedding for the target domain into an appropriate latent code for StyleGAN, thus yielding the desired target image. Comprehensive experimentation illustrates UniTranslator's consistent ability to generate highly plausible translation outcomes across diverse visual domains, irrespective of the extent of the domain gap. Moreover, our method exhibits exceptional performance in various real-world applications, including style mixing, stylization, and robust handling of translations even under degraded conditions.

The contributions of this work are:



(a) Source Image (b) GP-UNIT [20] (c) DiFa [17] (d) DiffusionCLIP [18] (e) PULSE [22] (f) UniTranslator (**ours**)

Fig. 2. The first row to the last row illustrate visual domain transformations from Metfaces to FFHQ (adjacent domains), AFHQ-dog to FFHQ (far-off domains), and AFHQ-cat to LSUN-church (intensively far-off domains). While GP-UNIT (b) can convert source domain images (a), it suffers from inadequate cross-domain correspondences and compromised image quality. Few-shot (c) or diffusion-based (d) domain adaptation methods display sensitivity to the magnitude of the domain gap. Even in the case of adjacent domains, these methods only result in minor changes to the input image towards the target domain. PULSE (e), lacking decoupling strategies, leaves remnants of source domain patterns when confronted with significant domain gaps. In contrast, UniTranslator (f) consistently achieves high-quality image transformations while upholding domain correspondence despite substantial visual disparities between the domains.

- We introduce UniTranslator for universal visual domain translation. By harnessing the domain-neutral capabilities of CLIP as a bridging conduit, UniTranslator empowers seamless conversions from any real-world source domain to a specified target domain, all accomplished with a single source image.
- We propose a decoupling module to efficiently extract cross-domain correspondences and integrate them with target-specific semantics to optimize the expected CLIP embedding.
- We design a specialized CLIP2P mapper that connects the CLIP and StyleGAN's spaces. This connection allows us to leverage the powerful generative capability of StyleGAN, leading to high-quality results.
- Extensive experiments demonstrate that UniTranslator performs favorably against state-of-the-art models regarding image quality, visual correspondences, and diversity. Furthermore, our method performs effectively in various real-world applications, including style mixing, stylization, and translations, even in challenging scenarios.

2 Related Work

Unsupervised Image-to-Image translation. Numerous unsupervised image-to-image translation methods [4], [6], [9], [10], [11], [12], [14], [15], [24], [25] have been developed to transfer images from the source to target domains. Zhu *et al*. [24] propose the cycle consistency constraints to learn the mapping between the source and target domains without paired data. In [15], Baek *et al*.introduce a guiding network for fully unsupervised scenarios. Choi *et al*. [9], [10] develop StarGAN, a model integrating multiple mappings, yet it requires data from multiple domains and cannot handle unseen inputs. On the other hand, Liu *et al*. [14] present FUNIT for few-shot translation to unseen classes.

Notably, these methods do not adapt well to situations where the target domain significantly differs from the source domain. GP-UNIT [20] specifically focuses on constructing pose mappings between complex domains with notable visual discrepancies. However, it is less effective in dealing with domains beyond the scope of ImageNet or intensely far-off domains. In contrast, our UniTranslator, as a hybrid technique, relies solely on a single reference image from the source domain during training, thereby circumventing limitations imposed by training data. Furthermore, it maintains high-level visual continuity across visual realms, rendering it highly suitable for translating between significantly distant domains.

Few shot/text-driven domain adaptation. A plethora of few-shot/text-driven domain adaptation methods have recently been proposed [16], [17], [26], [27], [28], [29], [30], [31] to train a generator using limited examples or text prompts. Ojha *et al*. [16] propose fine-tuning a pre-trained source generator using 10 target images to preserve relative similarities and differences in the source domain while avoiding mode collapse and Xiao *et al*. [26] consider spatial structural alignment between domains to enhance generative

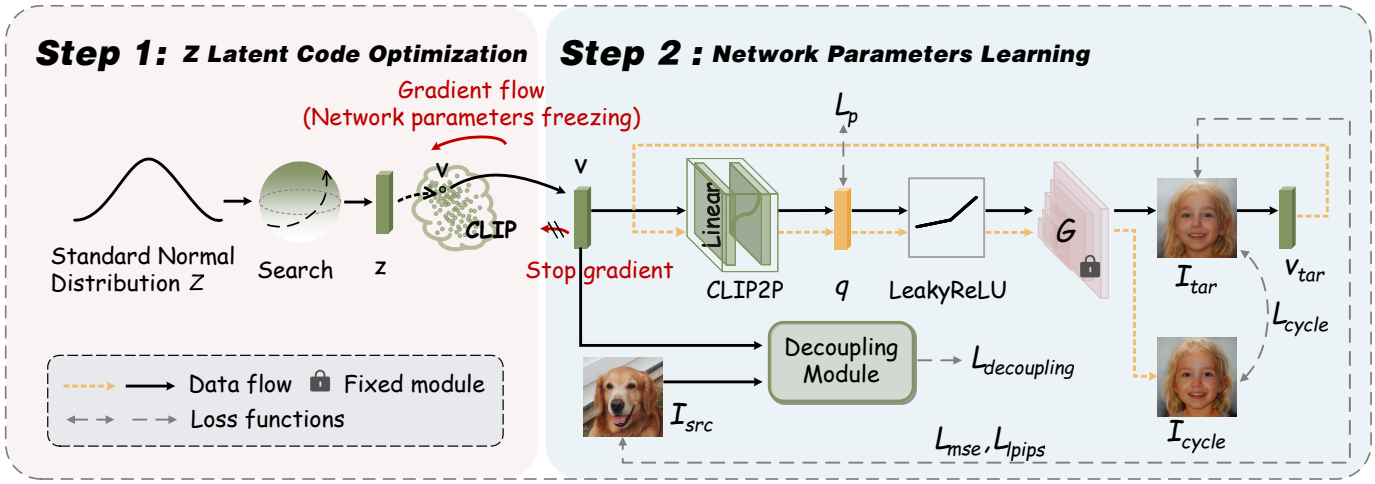


Fig. 3. Overview of UniTranslator. It leverages the decoupling module to extract domain-agnostic semantics and integrates them with target-specific information, resulting in a refined CLIP embedding with robust cross-domain correlations. This enhanced CLIP embedding will more effectively guide the search for an optimal z code. Moreover, the CLIP2P mapper is engineered to map the CLIP embedding into P space, reducing the likelihood of it falling outside of StyleGAN's latent space. A demo video is included in the supplementary material.

quality. In [27], Gal *et al.*introduce a text-guided adaptation method that constructs a directional CLIP loss based on a collinear relationship, and Zhu *et al.* [28] further improve this relationship by employing a GAN inversion method to determine the domain-gap direction. Recently, Zhang *et al.* [17] develop an attentive style loss and a selective generation strategy to promote the diverse generation and faithful adaptation in a one-shot scenario.

Despite the significant progress achieved by these methods, their effectiveness might be limited when dealing with substantial inter-domain gaps. In contrast, our approach directly explores the target manifold, ensuring that outcomes reside within the target space while maintaining cross-domain correspondences.

GAN Inversion. GAN inversion aims to find a latent code capable of faithfully reconstructing a given real image using a pre-trained generator. These methods can be categorized as encoder-based [32], [33], [34], [35], [36] and optimization-based [37], [38], [39], [40]. The former directly trains an encoder to map real images to latent codes. For instance, Richardson *et al.* [32] and Xu *et al.* [33] concentrate on designing the encoder structure. Tov *et al.* [41] consider the properties of W and $W+$ space, enhancing the editability of reconstructed images, and Alaluf *et al.* [35] propose an iterative feedback mechanism to facilitate the learning process. In [36], Bai *et al.*utilize padding space to enrich the representation capacity of the latent space, thereby refining spatial details. However, learning-based methods typically require a large number of training images and might not be practical when training samples are limited. In contrast, optimization-based methods can infer a single image at a time. Kang *et al.* [34] jointly optimize the extended f and $w+$ latent codes for faithful reconstruction of out-of-range and unaligned real images, and Abdal *et al.* [39] propose conditional exploration using continuous normalizing flows. Recently, Xu *et al.* [40] introduce consecutive images to strike a balance between editability and fidelity.

Motivated by the performance of the optimization-based inversion methods, we propose a new translation paradigm named UniTranslator. This approach entails seeking optimal latent codes guided by our tailored objective.

3 Method

3.1 Overview

The overview of our method is depicted in Fig. 3 and a demo video in the supplementary material. Given a single source image, UniTranslator aims to:

- Discover an optimal embedding corresponding to a target domain image while preserving cross-domain relationships with the source image.
- Map this embedding into StyleGAN's latent space.

We begin by navigating the Z space, employing a dual-branch architecture to achieve both goals through a hybrid learning approach. One branch is the decoupling module. Specifically, we utilize the source image as a reference and leverage the domain-neutral capabilities of the CLIP space, which aligns images with prompts of neutral classes. This alignment enables us to extract domain-agnostic information, represented as abstract CLIP embeddings. We achieve this by modeling the relationships among various combinations of embedding components. Merging domain-agnostic semantics with target-specific information guides the optimization process towards a more suitable embedding.

The other branch enhances the quality and diversity of image translation by tapping into the impressive generative capability of StyleGAN. To achieve this goal, we introduce a CLIP2P mapper as a crucial link between the CLIP space and StyleGAN's P space [42] (the deactivated space of W space). Note that directly converting CLIP embeddings to W space using only a single input image presents significant challenges. Mapping accurately to this complex latent distribution requires designing specific network modules or objectives, which is difficult with limited input data. Instead, we opt to use the P space as an intermediary, leveraging its properties to effectively transfer the desired CLIP embedding into StyleGAN's native latent space, resulting in high-quality output generation. Furthermore, we

do not consider $W+$ space due to its higher degrees of freedom, which complicates optimization when limited cues are available from a single image. This increased flexibility could lead to deviations from the StyleGAN target manifold, undermining the quality of the generated images.

Our UniTranslator operates through a two-step process during each training iteration. We optimize the latent code z in the first step while keeping the network parameters frozen. This step leverages the valuable information stored in the network parameters to guide the search for the optimal z code. In the second step, we fix the discovered z code and update the network parameters.

3.2 Decoupling Module

The main goal of the proposed decoupling module is to use CLIP's image-text alignment capability to extract domain-agnostic information. We initially convert the z code into a CLIP embedding. Following the implicit assumption from Corgi [43] that each neutral domain conforms to a high-dimensional Gaussian distribution, we begin by generating 5,000 target images corresponding to the selected target domain. These images are then processed through CLIP image encoder E_I to acquire their respective embeddings. Subsequently, We calculate statistics for these embeddings, including the standard deviation σ_{CLIP} and the mean μ_{CLIP}. Using these statistics, we transform the z code as an embedding v within the CLIP space:

$$v = \sigma_{CLIP} z + \mu_{CLIP}. \tag{1}$$

Fig. 4 shows a detailed illustration of the module structure. Regarding the module inputs, prompt templates are also required apart from the CLIP embedding v and the provided source domain image I_{src}. These inputs are split and fed into two separate streams based on whether they pertain to the source or target domain. Each stream contains two Multi-Layer Perceptrons (MLPs), with shared parameters across streams.

For the source domain stream (lower part of Fig. 4), an embedding $m_i^{src} \in \mathbb{R}^{512\times 1}$ of the source image is generated by the CLIP image encoder. This embedding is passed through the first MLP (MLP1), generating a 1024-dimensional vector. To extract domain-agnostic information, we divide this vector into two 512-dimensional embeddings f_s^{src} and f_a^{src} to learn domain-specific and domain-agnostic information. It is important to note that these two types of information are expected to be independent.

A similar process is applied to the target domain stream. The main difference is that the input to MLP1 is the CLIP embedding v, which is for inverting the target image. As a result, the target domain stream also yields two corresponding embeddings f_s^{tar} and f_a^{tar}. The first loss function $\mathcal{L}_o$ is:

$$\mathcal{L}_o = \frac{f_s^{tar} \cdot f_a^{tar}}{|f_s^{tar}||f_a^{tar}|} + \frac{f_s^{src} \cdot f_a^{src}}{|f_s^{src}||f_a^{src}|}. \tag{2}$$

Minimizing $\mathcal{L}_o$ enforces that the domain-specific embeddings and their corresponding domain-agnostic embeddings become orthogonal, guaranteeing their independence.

Furthermore, we align f_s^{tar} and f_s^{src} with their respective text embeddings m_t^{tar} and m_t^{src} produced by CLIP's text encoder E_T to ensure that they genuinely carry domain-specific information. This alignment is accomplished using the following loss function $\mathcal{L}_s$:

$$\mathcal{L}_s = (1 - \frac{m_t^{tar} \cdot f_s^{tar}}{|m_t^{tar}||f_s^{tar}|}) + (1 - \frac{m_t^{src} \cdot f_s^{src}}{|m_t^{src}||f_s^{src}|}). \tag{3}$$

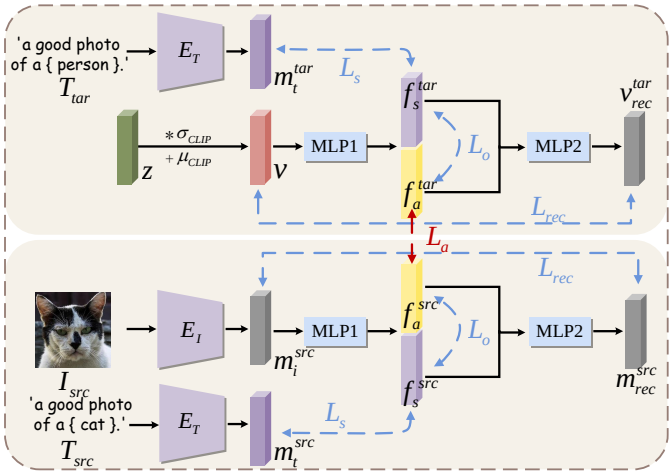


Fig. 4. Illustration of the proposed decoupling module.

Similar to the approach by Radford et al. [23], we utilize 80 diverse sentence templates, such as 'a photo of the {}', 'a photo of my {}', and 'a good photo of a {}'. The resulting embeddings from these prompts are averaged separately for the source and target domains. We have templates with the neutral class for both domains. With the impact of the above two loss functions, the remaining embeddings f_a^{tar} and f_a^{src} inherently capture domain-agnostic information. Moreover, we employ $\mathcal{L}_a$ to constrain their similarity, thereby aiding in the extraction of cross-domain correspondences:

$$\mathcal{L}_a = ||f_a^{tar} - f_a^{src}||_1. \tag{4}$$

We now create a combined embedding comprising f_s^{tar} and f_a^{tar}, encompassing target-specific and domain-agnostic information. This concatenated embedding is then fed through the second MLP (MLP2) for reconstructing a 512-dimensional vector v_{rec}^{tar}. The last loss term is the reconstruction loss $\mathcal{L}_{rec}$. This process can be formulated as:

$$\begin{aligned} &v_{rec}^{tar} = \text{MLP2}(f_s^{tar} \oplus f_a^{tar}), m_{rec}^{src} = \text{MLP2}(f_s^{src} \oplus f_a^{src}), \\ &\mathcal{L}_{rec} = ||v - v_{rec}^{tar}||_1 + ||m_i^{src} - m_{rec}^{src}||_1, \end{aligned} \tag{5}$$

where $\oplus$ denotes concatenation operation. This loss maintains consistency between the reconstructed vectors and their corresponding image embeddings throughout the network parameters learning step. While optimizing the z code, we aim to use the embedding v_{rec}^{tar}, which incorporates both target-specific and domain-agnostic information, to guide the search for an enhanced z.

Finally, the learning objective of the decoupling module is formulated as:

$$\mathcal{L}_{decoupling} = \mathcal{L}_o + \mathcal{L}_s + \mathcal{L}_a + \mathcal{L}_{rec}. \tag{6}$$

Previous latent code optimization techniques [44], [45] aim to enhance reconstruction fidelity or editability within StyleGAN's domain or a slightly regularized version to

TABLE 1
The KL divergence between the true latent distribution in P space and the Gaussian distribution constructed using the statistics of the latent codes sampled from the P space for various target domains.

Target Domain	KL Divergence (P (true) w.r.t P (pseudo))	KL Divergence (P(pseudo) w.r.t P (true))
FFHQ	0.6883	0.7016
Anime	1.0853	1.0964
E621Faces	1.1810	1.2037
LSUN-church	6.3166	6.3641

mitigate out-of-domain issues. In contrast, our decoupling module utilizes CLIP's domain-neutral features to maintain essential cross-domain correspondences. This approach liberates the solution space from being restricted to a single domain, allowing us to model relationships between domain-specific and domain-agnostic features across different domains and modalities. Our specialized objective thus establishes robust cross-domain correspondences.

3.3 CLIP2P Mapper

We analyze another branch within the framework where the CLIP2P mapper is situated. First, we introduce the concept of the P space, defined by II2S [42] as $p = \text{LeakyReLU}_{5.0}(w)$, where $p \in P$ and $w \in W$ represent two latent codes sampled from their respective spaces. The W space can be regarded as the activated counterpart of the P space. Assuming a Gaussian distribution for the P space, such a relationship can be formulated (as adopted by the official code of PULSE):

$$w = \text{LeakyReLU}_{0.2}(\sigma_P z + \mu_P), \tag{7}$$

where σ_P and μ_P indicate the standard deviation and mean of the Gaussian distribution. With this assumption, we can use the statistics of sampled CLIP embeddings to construct a Gaussian distribution (refer to Eq. (1)) and learn the latent distribution in P space through a linear layer that preserves the distribution type, yielding the w latent code as:

$$w = \text{LeakyReLU}_{0.2}(\text{Linear}(v)). \tag{8}$$

However, as II2S indicates, the latent distribution in P space is essentially an unimodal distribution resembling a Gaussian but not a true Gaussian distribution. To analyze the deviation from this assumption in transformations across different target domains, we sample 5000 latent codes from the P space of each target domain and calculated the Kullback-Leibler (KL) divergence between the Gaussian distribution established using the statistics of these codes (referred to as P (pseudo)) and the true distribution of these codes (refer to as P (true)). As shown in Table 1, there is indeed a disparity between the distributions P (true) and P (pseudo), which varies across different target domains. Thus, the w code obtained via Eq. (8) may lead to unsuccessful transformations in specific target domains (*e.g.*, E621Faces, LSUN-church), as it may fall outside the true latent space of StyleGAN.

We propose adding a nonlinear function $\mathcal{M}(\cdot)$ after the linear layer to learn the unimodal distribution in P space to remedy this. It is defined as:

$$\mathcal{M}(x) = \begin{cases} e^{h(x-\mu)} - 1, & x > \mu, \\ -e^{-j(x-\mu)} + 1, & x \leq \mu. \end{cases} \tag{9}$$

Here, h, j, μ are all learnable parameters, and e is Euler's number. The reasons for this function are threefold. First, while we introduce nonlinearity to ensure that the resulting distribution deviates from a Gaussian shape, it is crucial to maintain the function monotonicity. This guarantees the Gaussian distribution generated by the added linear layer remains unimodal after undergoing such a nonlinear transformation. Second, we introduce three degrees of freedom through parameters h, j, and μ. Among these, μ controls the peak position of the resulting unimodal distribution. Meanwhile, h and j control the curvatures on either side of the peak, ensuring the asymmetric distribution. Third, to ensure the continuity of this function, biases of -1 and $+1$ are applied at different intervals.

The CLIP2P mapper denoted as CLIP2P$(\cdot)$, comprises a linear layer and the above nonlinear mapping function. It is effectively trained using the distribution P (pseudo) as supervision through the loss term $\mathcal{L}_g$:

$$\mathcal{L}_g = ||\sigma_P z + \mu_P - q||_1, \tag{10}$$

where $q = \text{CLIP2P}(v) = \mathcal{M}(\text{Linear}(v))$ indicates the output after applying the CLIP2P mapper. We note the inclusion of $\mathcal{L}_g$ does not alter the output distribution type from the CLIP2P mapper. Instead, it enhances training stability.

We examine the KL divergence between the true latent distributions in CLIP and P spaces across different target domains. For each target domain, we sample 5000 latent codes. Table 2 shows that the disparity between these two distributions varies across domains. Consequently, adjusting the hyperparameter related to $\mathcal{L}_g$ based on the target domain is necessary although challenging. We set the hyperparameter λ_p of $\mathcal{L}_g$ as a learnable parameter to address this. This enables us to formulate the final loss function $\mathcal{L}_p$ for the CLIP2P mapper as:

$$\mathcal{L}_p = \text{ReLU}(\lambda_p)\mathcal{L}_g. \tag{11}$$

In practice, we observe that λ_p sometimes converges to 0 (*e.g.*, when using FFHQ as the target domain). This suggests that effective learning can occur with less reliance on the supervision from P (pseudo). However, in other scenarios, such as when working with target domains like Anime or LSUN-church, supervision is necessary to constrain parameters learning, and λ_p dynamically adjusts to an appropriate non-zero value based on the target domain.

While techniques such as StyleCLIP's latent mapper [46] also perform latent code mapping, they focus on in-domain mappings within StyleGAN's latent space, which are not designed for cross-domain translation tasks that involve significant domain gaps, such as converting a painting into a photo [27]. Our CLIP2P mapper, on the other hand, bridges the CLIP and P spaces by leveraging their unique characteristics. This ensures that the latent code transformed from the CLIP space accurately resides within StyleGAN's target manifold. By doing so, our approach allows for the search of target embeddings with cross-domain correspondences in the open-world CLIP space, while effectively utilizing StyleGAN's generative priors to achieve universal domain translation.

TABLE 2
The KL divergence between the true latent distributions in the P space and CLIP space for various target domains.

Target Domain	KL Divergence (P (true) w.r.t CLIP (true))	KL Divergence (CLIP (true) w.r.t P (true))
FFHQ	0.4754	0.4587
Anime	1.0571	1.0351
E621Faces	1.1687	1.1141
LSUN-church	3.6901	4.2885

3.4 Learning Objectives

Aside from the above-explained loss terms $\mathcal{L}_{decoupling}$ and $\mathcal{L}_p$, we use three other loss terms to train the UniTranslator: $\mathcal{L}_{mse}$, $\mathcal{L}_{lpips}$, and $\mathcal{L}_{cycle}$.

In addition to preserving semantic correlations across domains, we maintain visual aspects such as color tone and perceptual relationships in cross-domain translation tasks. As observed in DiffuseIT [19] and Zhu *et al*. [28], relying solely on CLIP-based semantic alignment is insufficient to ensure color consistency between the input and output images. To address this, we incorporate a loss term $\mathcal{L}_{mse}$, similar to its use in DiffuseIT, to ensure that color matching is maintained before and after translation. This loss enforces a constraint on the Euclidean distance between the source image I_{src} and the target image I_{tar}, formulated as:

$$\mathcal{L}_{mse} = ||I_{tar} - I_{src}||_2, \tag{12}$$

and

$$I_{tar} = G(\text{LeakyReLU}_{0.2}(q)), \tag{13}$$

where $G(\cdot)$ denotes the pretrained StyleGAN generator. It is important to note that our approach mitigates the risk of image blurriness by generating images through latent space traversal rather than the traditional feature space-to-output method. With StyleGAN's robust generative capabilities, as long as the latent code remains within the StyleGAN target manifold, constrained by our CLIP2P mapper and $\mathcal{L}_p$ loss, it produces sharp and high-quality images.

The other term, $\mathcal{L}_{lpips}$ [47], leverages deep features to guide the perceptual relationship between I_{tar} and I_{src}:

$$\mathcal{L}_{lpips} = ||F(I_{tar}) - F(I_{src})||_2, \tag{14}$$

where $F(\cdot)$ represents the perceptual feature extractor.

The last loss term is a cycle loss, denoted as $\mathcal{L}_{cycle}$. In UniTranslator, after generating the translated result I_{tar}, we proceed to feed it into CLIP's image encoder, obtaining a CLIP embedding $v_{tar} = E_I(I_{tar})$. Subsequently, we pass v_{tar} as input to the CLIP2P mapper and replicate the remaining process. As such, we have a new output, denoted as I_{cycle}. We then impose the constraint $\mathcal{L}_{cycle}$ to ensure consistency between the two images, I_{tar} and I_{cycle}. This secures that the CLIP2P mapper only transforms the space type (from CLIP to P) without modifying the image semantics:

$$\begin{aligned} I_{cycle} &= G(\text{LeakyReLU}_{0.2}(\text{CLIP2P}(v_{tar}))), \\ \mathcal{L}_{cycle} &= ||I_{tar} - I_{cycle}||_2. \end{aligned} \tag{15}$$

Note that although our initial assumption regarding the distribution in CLIP's neutral domain as a Gaussian may not strictly hold, the cycle loss can mitigate the risks arising from this violation by constraining the latent vector to carry more semantic information from the CLIP space.

Finally, our total loss function $\mathcal{L}_{total}$ is formulated as:

$$\begin{aligned} \mathcal{L}_{total} =& \lambda_{mse}\mathcal{L}_{mse} + \mathcal{L}_{lpips} + \mathcal{L}_{decoupling} \\ &+ \mathcal{L}_{cycle} + \mathcal{L}_p, \end{aligned} \tag{16}$$

where λ_{mse} denotes a hyperparamter for balancing loss terms.

ALGORITHM 1: UniTranslator

Data: Initial latent code $z_{(0)}$, hyperparameter $\lambda_p^{(0)}$, and learnable network parameters $\theta_{(0)}$; Statistics σ_{CLIP}, μ_{CLIP}, σ_P and μ_P; Source Image I_{src}; Prompt templates; Hyperparameters $\{\lambda_i\}$; Iteration number N; Pretrained StyleGAN and CLIP models
Result: Optimal latent code $z_{(N)}$ and Target Image I_{tar}

```
for i ← 1 to N do
    while Step 1: z latent code optimization do
        Update the latent code z_(i) with z_(i−1), θ_(i−1), and λ_p^(i−1) by Eq. (16);
    end
    while Step 2: Network parameters learning do
        Update the learnable parameters θ_(i) and λ_p^(i) with z_(i) by Eq. (16);
    end
end
Obtain I_tar with z_(N) by Eq. (13);
```

4 Experiments

4.1 Implementation Details

To implement UniTranslator, we use a hybrid learning scheme and outline the workflow in Algorithm 1. In each iteration, we start with step 1, optimizing the z code using a spherical optimizer [22] with a learning rate of 0.4. Then, we move to step 2, where we focus on learning the network parameters using the ADAM optimizer [48] with exponential decay rates of $\beta_1 = 0$ and $\beta_2 = 0.99$. The learning rate for updating all parameters is set to 0.002, except λ_p, which is assigned a higher value of 0.1. The hyperparameter λ_{mse} in our objective is set to 10. Approximately 35 iterations are conducted for each source image, taking between 20 and 45 seconds on a PC with an Nvidia GeForce RTX 3090.

4.2 Evaluated Methods

We compare UniTranslator with numerous state-of-the-art methods based on learning (*e.g.*, VQ-I2I [5], StarGAN2 [10], and GP-UNIT [20]), one-shot domain adaptation (*e.g.*, DiFa [17]), inference (*e.g.*, PULSE [22]), and diffusion models (*e.g.*, DiffusionCLIP [18] and DiffuseIT [19]).

It is worth noting that learning-based methods require a substantial amount of training data from both source and target domains. The domain adaptation method, *e.g.*, DiFa, requires a target image to fine-tune its pre-trained source generator. Diffusion-based methods utilize two text prompts: one for the source domain and another for the target domain. These prompts are used to fine-tune the pre-trained diffusion model or govern the sampling process. On the other hand, inference-based methods such as PULSE and

UniTranslator perform direct inference based on the input image (and prompt templates). The settings and hyperparameters of the evaluated methods are according to their official source codes.

4.3 Datasets

We use eight source-to-target translation mappings to assess the universality of all the comparisons. These mappings cover a wide range of domain heterogeneity, including cases where domains are 'adjacent', such as Metfaces [49]→FFHQ [50] and AFHQ-cat [10]→E621Faces [51], as well as cases where domains are 'far-off', such as AFHQ-cat→Anime [52], AFHQ-cat→FFHQ, AFHQ-dog→FFHQ, and AFHQ-wild→FFHQ. To provide a more challenging evaluation, we also include mappings between 'intensely far-off' domains, such as LSUN-Church [53]→FFHQ and AFHQ-cat→LSUN-Church. Note that we do not address intra-domain translations, such as male-to-female mapping within the FFHQ domain, as these can be easily achieved with state-of-the-art editing methods [44], [54]. Instead, our focus is on cross-domain translations, even those involving minimal domain gaps, such as the adjacent domains we target.

For datasets with predefined train-test splits, such as AFHQ and LSUN-church, we use the default training set to train learning-based competitors. In cases where predefined splits are unavailable, as with MetFaces, FFHQ, E621Faces, and Anime, we perform a random 7:3 split between the training and test sets. For fine-tuning the domain adaptation method DiFa, we randomly select a single training image from the target domain. Note that inference-based and diffusion-based methods conduct inference directly on the test set or utilize text prompts without relying on the training images. All reported performances are based on results obtained from the test set.

The datasets used in the evaluations contain images of different resolutions: 1024×1024 (*e.g.*, FFHQ, Metfaces), 512×512 (*e.g.*, AFHQ, Anime, and E621Faces), and 256×256 (*e.g.*, LSUN-church) pixels. Note that VQ-I2I, GP-UNIT, StarGAN2, DiffusionCLIP, and DiffuseIT are limited to generating results at a resolution of 256×256. For DiFa, the generative resolutions align with those of the source domain datasets, as they depend on the source generators. In contrast, PULSE and UniTranslator consistently produce results at the same resolution as the target domain dataset.

4.4 Evaluation Metrics

We use various metrics to analyze the quality of the generated results and the perceptual correspondence between inputs and outputs. Specifically, we use the no-reference metric Naturalness Image Quality Evaluator (NIQE) [55] to assess image quality, with a particular focus on the perceived realism of the results, including any potential distortions or artifacts. Additionally, in cross-domain translation tasks, the generated images must bear perceptual similarity to the corresponding reference images. We utilize the LPIPS metric [47], which relies on deep features. It is calculated for each input-output pair, with an average score reported.

Furthermore, after producing the transformed images for each translation task, we measure the similarity between their CLIP embeddings and those of the target dataset to assess how well the output images fit into the target domain. We also include a user study to support our evaluations, both of which are detailed in the supplementary materials.

It is important to note that while the Fréchet Inception Distance (FID) [56] and the Inception Score (IS) [57] are widely used to evaluate image quality, they are not ideally suited for cross-domain translation tasks. A detailed discussion of these metric choices is provided in the supplementary materials.

4.5 Qualitative Evaluation

Qualitative evaluation results, as shown in Fig. 5 and Fig. 6, demonstrate the limitations of learning-based methods (VQ-I2I, StarGAN2, and GP-UNIT) in achieving high-quality translations. While these methods succeed in translating certain patterns into target domains, the synthesized images are not less appealing. This issue primarily arises because their feature extraction mechanisms excel mainly within visually similar domains. Specifically, VQ-I2I and StarGAN2 overlook high-level correspondences between domains, restricting their capability to handle translations across visually distinct domains. GP-UNIT, which leverages the generative prior from BigGAN [21], can synthesize images in complex domains with substantial visual disparities. However, due to the inadequate disentanglement of BigGAN's latent space, GP-UNIT cannot effectively establish effective correspondences with the source domain concerning contours and color tones, resulting in less realistic outcomes. Additionally, GP-UNIT exhibits a lack of generalization when applied to domains beyond the training range of its pose encoder, further diminishing its performance. These limitations significantly undermine the universality of learning-based methods. In contrast, our UniTranslator performs well by requiring only a single source domain image, disregarding the source domain's range, and extracting domain-agnostic information through a decoupling module. It achieves robust correspondences with the source domain across various aspects, including poses, contours, and color tones. Furthermore, by harnessing the better-disentangled generative prior of StyleGAN, UniTranslator consistently generates highly realistic outputs.

The domain adaptation method, DiFa, and diffusion-based methods, DiffusionCLIP and DiffuseIT, do not translate the images well, as the outputs closely resemble the sources. These methods utilize CLIP losses applied to pre-trained source models for cross-domain translations. However, the qualitative results underscore the insufficiency of this approach in bridging substantial domain gaps. In contrast, our approach uses the CLIP2P mapper to navigate the StyleGAN target manifold, ensuring that the output results reside within the target domain.

It is worth noting that although PULSE also leverages the generative prior of StyleGAN, it often produces ambiguous translations. Its guidance for traversal relies solely on pixel-wise constraints applied to downsampled images, a strategy well-suited for super-resolution methods but overly simplistic for cross-domain translations. On the other hand, our approach considers both the establishment of cross-domain correspondences and the quality of the generated results,

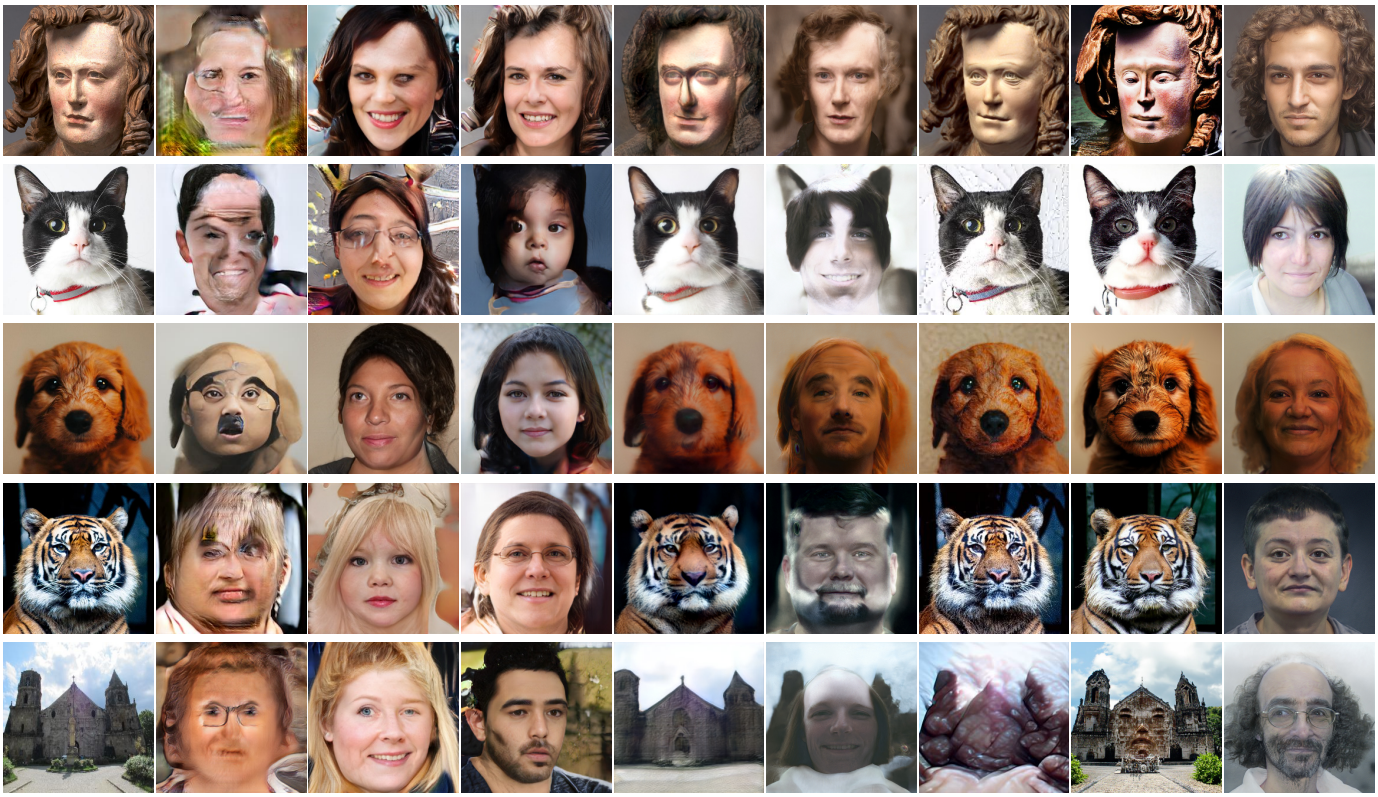


Fig. 5. Qualitative comparison of our UniTranslator with state-of-the-art methods for translating $\mathcal{X}$ to FFHQ. Note that significant issues encountered by other methods, including severe distortions (VQ-I2I), poor cross-domain correspondences (GP-UNIT and StarGAN2), presence of source domain patterns (VQ-I2I and PULSE), and even the inability to generate target domain patterns (Difa, DiffusionCLIP, and DiffuseIT).

thereby translating images effectively. Additionally, due to UniTranslator's minimal increase in learnable parameters, it steadily translates images through optimization without excessive overheads.

4.6 Diversity

The correspondences become increasingly abstract and less rigid as the gap widens between the source and target domains. This natural relaxation also affects the accompanying target-specific information, rendering it similarly more flexible. The confluence of these two types of information can lead to multiple valid solutions, as the inherent multiplicity of reasonable outputs for a given input in cross-domain translation. From the visual perspective, the rationality of the results must take into account cross-domain correspondences, while diversity stems from the flexibility in both cross-domain correspondences and target-specific information.

Fig. 7 shows comparisons that encapsulate this aspect of diversity. Our approach generates diverse yet reasonable results through multiple inferences. The results show that the extracted domain-agnostic information possesses flexibility without sacrificing reasonableness, in line with the fact that strict cross-domain correspondences may not always be guaranteed (such as when inferring a person's facial age based on an animal's face). In contrast, other methods often generate low-quality outputs with poor alignments with the source images or translated images with minimal variation.

4.7 Quantitative evaluation

Table 3 shows the NIQE and LPIPS scores of the evaluation results. Our method consistently performs favorably across various configurations and mapping tasks. These results demonstrate our method's ability to translate high-quality images reliably while preserving strong perceptual correlations between the sources and targets. We do not include the LPIPS score for DiFa and DiffusionCLIP, as their outputs predominantly remain within the source domain. As evident in Fig. 5 and Fig. 6, their translated outputs resemble the input images. Thus, assessing the perceptual correspondences of such results within the scope of cross-domain translation tasks becomes inherently inconsequential. Furthermore, we compute the similarity of CLIP embeddings between the generated results and the target dataset for each mapping task, assessing the degree to which the outcomes belong to the target domain. More details are provided in the supplementary material.

4.8 Ablation Study

Effectiveness of CLIP2P Mapper. To evaluate the effect of our CLIP2P mapper, we conduct a qualitative ablation study by removing the nonlinear mapping $\mathcal{M}(\cdot)$ and using the output of the linear layer as the q code in Eq. (10). The resulting images are displayed in the 7th column of Fig. 8. Removing this function noticeably degrades image quality in the E621Faces and LSUN-church domains, while

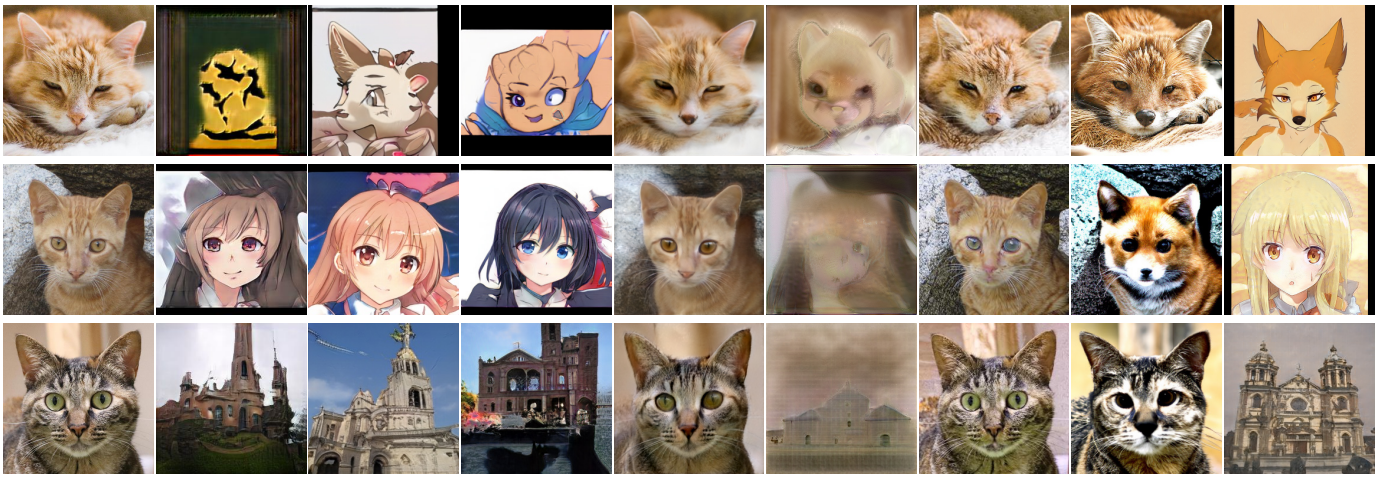


Fig. 6. Comparison with the competitors. First row: AFHQ-cat→E621Faces; Second row: AFHQ-cat→ Anime;Third row:AFHQ-cat→LSUN-church.

Fig. 7. Diversity comparison. Given one input, UniTranslator generates diverse and sensible results through multiple optimizations, while alternative methods often yield single-modal or lower-quality outputs.

TABLE 3
Quantitative comparison of our UniTranslator with state-of-the-art methods. We use NIQE and LPIPS Scores to assess the quality and perceptual similarity of the generated images. Both metrics are the lower, the better, and the best results are highlighted in bold with underline.

Type	Mapping	VQ-I2I		GP-UNIT		StarGAN2		DiFa		PULSE		DiffusionCLIP		DiffuseIT		**UniTranslator**	
		NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS	NIQE	LPIPS
Adjacent	Metfaces→FFHQ	4.81	0.68	4.44	0.62	4.47	0.60	7.47	-	5.07	0.50	5.65	-	5.79	0.54	**4.25**	**0.18**
Adjacent	AFHQ-cat→E621Faces	7.09	0.80	6.49	0.76	5.91	0.75	5.10	-	7.41	0.64	**3.75**	-	6.12	0.45	4.94	**0.26**
Far-off	AFHQ-cat→Anime	6.06	0.74	5.80	0.75	6.02	0.75	4.87	-	5.03	0.63	**4.07**	-	5.59	0.44	4.25	**0.26**
Far-off	AFHQ-cat→FFHQ	4.44	0.72	4.57	0.76	4.70	0.72	4.83	-	4.72	0.61	4.89	-	6.74	0.45	**3.69**	**0.23**
Far-off	AFHQ-dog→FFHQ	4.83	0.66	4.68	0.76	4.78	0.68	6.37	-	4.81	0.58	4.47	-	5.88	0.41	**4.12**	**0.22**
Far-off	AFHQ-wild→FFHQ	4.23	0.72	4.40	0.76	4.64	0.72	5.18	-	4.63	0.64	5.86	-	5.98	0.45	**4.17**	**0.23**
Intensely far-off	LSUN-church→FFHQ	5.11	0.80	4.43	0.81	4.47	0.81	5.69	-	4.82	0.65	5.86	-	6.09	0.45	**3.98**	**0.23**
Intensely far-off	AFHQ-cat→LSUN-church	4.69	0.72	4.23	0.76	4.57	0.76	4.40	-	6.01	0.67	4.33	-	5.69	0.44	**4.12**	**0.26**
/	Average	5.16	0.73	4.88	0.75	4.95	0.72	5.49	-	5.31	0.62	4.86	-	5.99	0.45	**4.19**	**0.23**

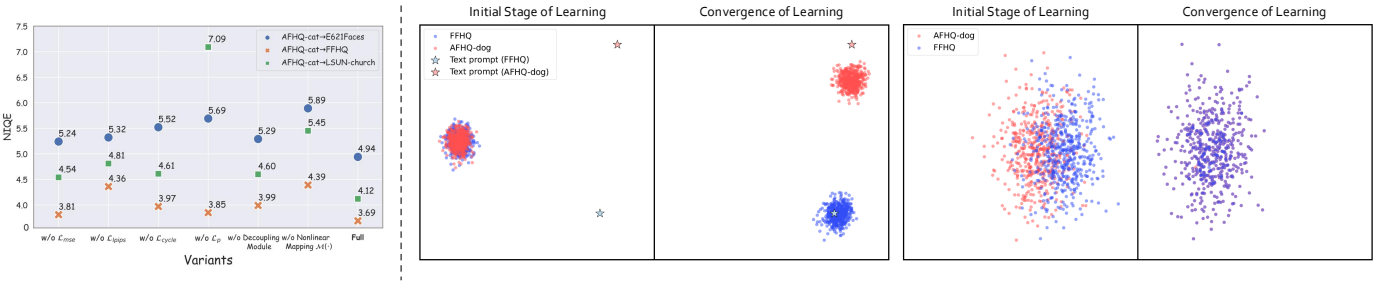
the effect on image quality in the FFHQ domain is relatively minor.

These results can be attributed to the differences between the distributions P (true) and P (pseudo) across various target domains, as observed in Table 1. Domains like FFHQ, where the corresponding P (pseudo) and P (true) distributions are closer, only require an additional linear layer to enable the q code to approximate the p code, even without the nonlinear mapping. Conversely, when the distribution P (pseudo) significantly deviates from the distribution P (true), learning the q code from the clip embedding v becomes more complex, necessitating our nonlinear mapping. Notably, even in the FFHQ domain, using nonlinear mapping can yield improved visual results, as it can more accurately approximate the distribution P (true).

Furthermore, we remove the linear layer from the CLIP2P mapper while keeping the nonlinear mapping for analysis. However, in practice, we encounter abnormally high loss values during the early iterations, resulting in training instability. This suggests that the small number of additional parameters provided by the linear layer is necessary to fit the required mapping better. We exclude the results of completely removing the entire CLIP2P mapper due to apparent compatibility issues between the CLIP and P spaces.

In addition to the qualitative ablation study, we conduct

Fig. 8. Ablation studies on network modules (*i.e.*, "w/o Decoupling Module", "w/o Nonlinear Mapping $\mathcal{M}(\cdot)$") and loss function designs (*i.e.*, "w/o $\mathcal{L}_{mse}$", "w/o $\mathcal{L}_{lpips}$", "w/o $\mathcal{L}_{cycle}$" and "w/o $\mathcal{L}_{p}$"). Each of these components contributes to the final quality of the results. (First row: AFHQ-cat→E621Faces; Second row: AFHQ-cat→FFHQ; Third row: AFHQ-dog→FFHQ; Last row: AFHQ-cat→LSUN-church.)



(a) Quantitative evaluation of ablation study (b) Visualization of domain-specific information (c) Visualization of domain-agnostic information

Fig. 9. Quantitative evaluation of ablation study and visualized features generated by the decoupling module.

a quantitative analysis, as shown in Fig. 9 (a), comparing the proposed method with its variations across different types of mappings. The results demonstrate that the most significant improvement in translation quality comes from the integration of our novel mapping function, highlighting its pivotal role in enhancing the overall system performance. An exception is the extreme case without the $\mathcal{L}_p$ loss in the LSUN-church target domain, which will be discussed in the context of effectiveness of loss functions. All observations indicate that our CLIP2P mapper significantly expands the boundaries of cross-domain translation universality.

Effectiveness of Decoupling Module. To demonstrate the decoupling module's importance, we analyze this component removed, and the qualitative results are presented in the 6th column of Fig. 8. A comparison with our full-featured model reveals that the decoupling module effectively leverages semantic information to construct correlation mappings related to pose, coarse outlines, and specific elements of fine attributes.

For example, in the 2nd row of Fig. 8, our method generates a white scarf on the girl's neck, corresponding to the white patch on the cat's chest in the source image. In contrast, the images generated without the decoupling module lack such a vivid correspondence. In the 3rd row, while other configurations tend to produce long curls to match the dog's curly fur, the absence of the decoupling module results in short hair. Even in scenarios characterized by a significant gap between source and target domains, such as the translation from a cat to a church, the decoupling module plays a crucial role by successfully matching the cat's pointed ears with the spires of the church.

Beyond its significant advantage in maintaining cross-domain correspondences, the decoupling module also enhances translation quality across source and target domains with varying levels of heterogeneity, as shown in Fig. 9 (a). This improvement is likely attributed to the effective utilization of the natural statistical properties inherent in the source image.

We also visualize domain-specific and domain-agnostic information to understand the decoupling module's effect. Using the translation of AFHQ-dog to FFHQ as an example, we sample 500 images from the AFHQ-dog domain and translate them to the FFHQ domain. We apply PCA to reduce the dimensions of domain-specific and domain-agnostic features learned by the decoupling module during the initial and convergence stages of learning, visualizing the results in Fig. 9 (b) and (c). Here, red dots represent source domain features, while blue dots indicate target domain features. Additionally, in Fig. 9 (b), we mark the average embeddings (also PCA-reduced) of the source and target prompt templates as red and blue stars. It is demonstrated that as training progresses, domain-specific information progressively shifts towards its respective text prompts, eventually clustering around them. Meanwhile, domain-agnostic information converges from a scattered state. All of these outcomes provide compelling evidence for the effectiveness of the decoupling module.

Effectiveness of Loss Functions. We analyze the roles of each loss term in the proposed method. First, we remove $\mathcal{L}_p$ from our total objective. As depicted in Fig. 8, when the target domain is E621Faces or LSUN-church, removing $\mathcal{L}_p$ causes the generated results to deviate almost entirely from the target domain. When the target domain is FFHQ, the absence of $\mathcal{L}_p$ results in a degradation of image quality, although the generated results remain near the face domain.

Furthermore, we analyze the effect of excluding $\mathcal{L}_{cycle}$. As illustrated in Fig. 8, the absence of $\mathcal{L}_{cycle}$ affects the semantic correspondences between the source and target domains. In this scenario, step 2 may alter the semantics carried by the v code, not just its distribution. The conflict with the function of step 1 results in the ineffectiveness of the hybrid learning strategy, leading to suboptimal results.

Finally, we examine the effects of $\mathcal{L}_{mse}$ and $\mathcal{L}_{lpips}$. As shown in Fig. 8, the omission of $\mathcal{L}_{mse}$ influences the color space consistency with regard to the input image (*e.g.*, skin tones), while the exclusion of $\mathcal{L}_{lpips}$ leads to a failure in establishing perceptual relationships (*e.g.*, determining the scale of the church).

Referring to Fig. 9 (a), the quantitative impact of each loss term is evident, aligning with our qualitative ablation analysis. While all loss terms contribute to performance improvement, it is particularly noteworthy that in the LSUN-church target domain, the $\mathcal{L}_p$ loss assumes an even more critical role. As shown in Table 2, the gap between the distributions in P and CLIP spaces is more pronounced in the LSUN-church domain than in other target domains. Therefore, utilizing $\mathcal{L}_p$ is vital in such domains, guiding the CLIP embedding towards the p latent code. All the evidence underscores the importance of these components in our approach to enhancing the quality of cross-domain image translations.

Fig. 10. We take four input images from different source domains, transform them into the target FFHQ domain, and save the corresponding w latent codes. Subsequently, we conduct smooth interpolation between these latent codes. The images in the corners indicate the source images and the 6×6 images in the middle depict the interpolated results.

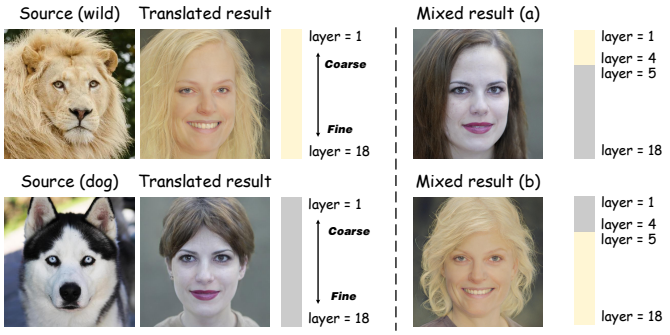


Fig. 11. Style mixing. Mixed result (a) is generated by injecting coarse wild style into the dog style, while mixed result (b) is produced by injecting fine wild style into the dog style.

5 Applications

One salient property of UniTranslator is its ability to derive in-domain w latent codes for its results, ensuring that controllability is not sacrificed for expressiveness. We perform smooth interpolation and style mixing experiments to demonstrate the quality of our latent codes. Furthermore, we conduct qualitative experiments involving style transfer and various degradation scenarios to demonstrate the robustness of our method.

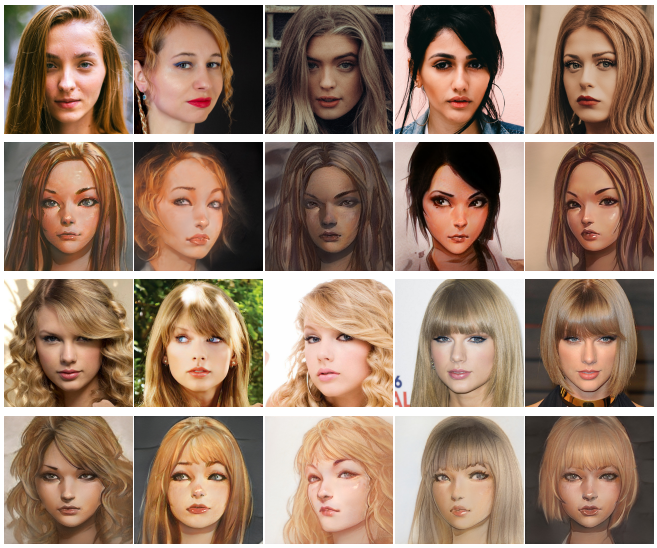


Fig. 12. Real images stylization. We transform real portraits into Ilya Kuvshinov's style. The first and third rows represent the real images, while the second and fourth rows showcase the stylized images.

5.1 Controllable Smooth Interpolation

The original W space of StyleGAN exhibits properties of smoothness and disentanglement. As a result, interpolation between two latent codes can manifest a smooth transition in multi-level attributes. Initially, we select four source images, each from AFHQ-cat, AFHQ-dog, AFHQ-wild, and LSUN-church, considering their diverse backgrounds, poses, and color tones. Utilizing the proposed method, we transform these images into the FFHQ domain, resulting in four distinct face latent codes. Interpolation between these latent codes seamlessly transitions multi-level attributes along either direction, as illustrated in Fig. 10. These results demonstrate that our approach can generate high-quality latent codes within the target domain, preserving the controllability of the W space.

5.2 Style Mixing

We use style mixing [34] to validate the quality of our latent codes, as shown in Fig. 11. Initially, we translate images of a lion and a dog into the FFHQ domain and save their respective w codes. Subsequently, we replicate and combine these latent codes. The first 4 layers of the synthesis network are injected with the lion's latent code, and the remaining 14 layers receive the dog's latent code, resulting in a mixed outcome (a). In this image, attributes such as pose and face shape are inherited from the lion, while the color scheme and finer details are drawn from the dog. For instance, the individual in result (a) features an angled pose, primarily derived from the lion's pose. However, the dark hair color and fair skin tone originate from the dog's dark head and white face.

Next, we reverse the latent codes of the lion and the dog to create the mixed result (b). Here, the lion influences the fine structure, and the coarse attributes are derived from the dog. The person in result (b) assumes a forward pose, adapted from the dog's pose, while their yellow skin tone is borrowed from the lion's features. These results demonstrate the capacity of the proposed method to achieve nuanced style mixing while preserving the integrity of the target images.

5.3 Stylization

Stylization, as one of the significant applications of image translation, warrants special attention. For the stylization experiments, we collect real portraits from Unsplash and Pexels websites and a set of portraits of the star Taylor Swift. These portraits are then translated into the artistic style of Ilya Kuvshinov, as shown in Fig. 12, yielding noteworthy results. The proposed method captures intricate details, such as the complex texture of hair colors in real scenarios. The results demonstrate the capability of our method to produce high-quality stylized images, even when dealing with smaller domain gaps.

5.4 Robustness in Degradation Scenarios

A translation application for real-world scenarios necessitates the ability to handle user-provided inputs. In such cases, the source domain is uncertain, and the image quality can also vary significantly. Inputs may include low-resolution or corrupted images, highlighting the importance of algorithmic robustness. Thus, we evaluate our method under two conditions: handling low-resolution inputs and corrupted inputs.

For low-resolution inputs, we experiment with 32×32 LR images as inputs and generate 1024×1024 HR results. PULSE can also generate results with a resolution of 1024 pixels. However, VQ-I2I, StarGAN2, GP-UNIT, and DiffusionCLIP cannot process LR images as inputs. As a result, we resize LR images to 256×256 and generate 256×256 HR results. DiFa, which fine-tunes the AFHQ generator, produces 512×512 images. These are presented in the first two rows of Fig. 13. VQ-I2I, GP-UNIT, and StarGAN2 cannot translate well with low-resolution images. DiFa and DiffusionCLIP manage only to generate blurry images. The results of PULSE retain some undesirable patterns from the source images, such as spots on a dog's face. In contrast, our method can generate high-resolution images, translating them to the target domain while maintaining high quality.

We also simulate corrupted images by randomly adding masks to the input images. Note that during this process, the constraints of $\mathcal{L}_{decoupling}$, $\mathcal{L}_{mse}$, and $\mathcal{L}_{lpips}$ are applied only to regions that are not masked. The results of these tests are presented in the last two rows of Fig. 13. Under these challenging conditions, the proposed method performs adaptively and stably, underscoring its potential to handle diverse and unpredictable real-world inputs.

6 Limitations

Our UniTranslator aims to transform images between visually distinct domains while maintaining domain correspondence. This translation framework can connect any real-world source domain to a chosen target domain. However, our approach is still limited by the generative capacity of StyleGAN2. Illustrative failure cases are depicted in Fig. 14. Certain expressions and poses, such as a dog sticking out its tongue or a cat curling up, are common in specific source

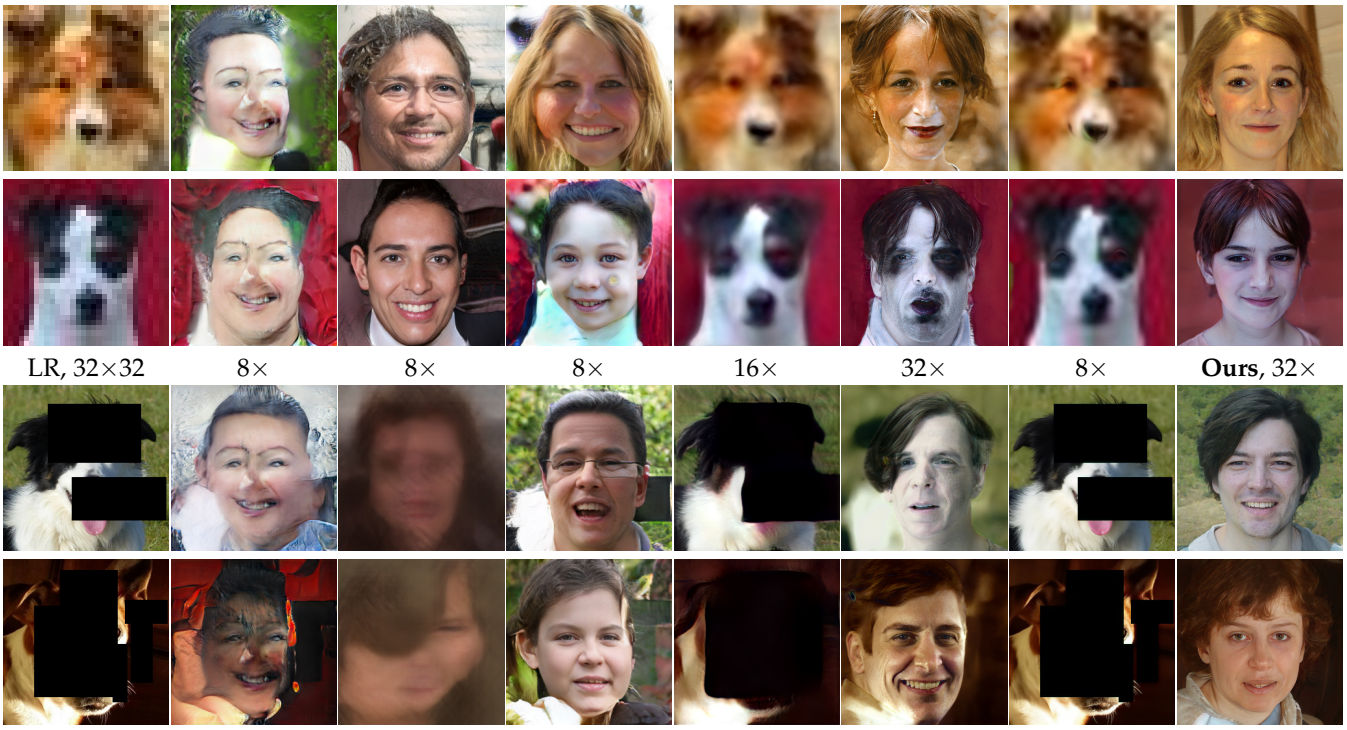


Fig. 13. A comparison of AFHQ-dog→FFHQ in degradation scenarios.

Fig. 14. Some failure cases of our method.

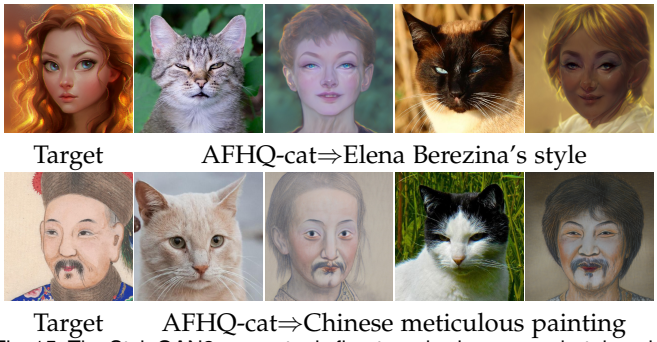


Fig. 15. The StyleGAN2 generator is fine-tuned using a one-shot domain adaptation method, DiFa, to obtain the representation of the unseen target domain.

domains, but their counterparts in target domains, such as humans, do not naturally assume these poses. Given that these rare expressions and poses were not sufficiently represented during the pre-training phase of StyleGAN2, generating such images remains challenging. Moving forward, our focus will pivot towards large-scale generators such as StyleGAN-XL [58] and GigaGAN [59], which demonstrate superior performance on weakly structured datasets, thereby potentially overcoming the current limitation.

The second limitation is that there may be a lack of large-scale data to train the target domain generator. Fortunately, few-shot domain adaptation methods provide a possible solution. By fine-tuning an off-the-shelf pre-trained generator using a few target samples, it is easy to get the target domain generator. As depicted in Fig. 15, we first pre-adapt the FFHQ generator by DiFa [17] using a sample in artist Elena Berezina's style or Chinese meticulous painting style. Subsequently, UniTranslator transforms a diverse range of source images into this new target domain.

In addition, our current method cannot handle inputs with multiple subjects, as StyleGAN exhibits limitations in generating images with multiple subjects. Furthermore, while the abstract nature of the cross-domain correspondences captured by the decoupling module provides flexibility in managing the scope of source domains, it compromises interpretability to a certain degree, which poses challenges in ensuring individual object correspondences in such cases. Addressing this limitation will be part of our future work.

7 Conclusion

In this work, we introduce UniTranslator, a pioneering paradigm that combines the domain-neutral capabilities of CLIP with the practical generative ability of StyleGAN for universal visual domain translation. Using the cutting-edge vision-language model CLIP, we develop a decoupling

module that extracts abstract and domain-agnostic semantics from CLIP representations. Furthermore, we introduce CLIP2P mapper, a non-linear mapping technique, to bridge CLIP and StyleGAN's latent spaces, effectively utilizing StyleGAN's generative priors. Extensive experimental results, both qualitative and quantitative, demonstrate that the proposed method performs favorably against state-of-the-art models regarding semantic correspondences and visual quality. Finally, we also demonstrate the versatility and robustness of UniTranslator through diverse applications.

REFERENCES

[1] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," in *ICML*, 2023. 1

[2] Z. Wang, Y. Li, X. Chen, S.-N. Lim, A. Torralba, H. Zhao, and S. Wang, "Detecting everything in the open world: Towards universal object detection," in *CVPR*, 2023, pp. 11 433–11 443. 1

[3] Q. Shen, X. Yang, and X. Wang, "Anything-3d: Towards single-view anything reconstruction in the wild," *arXiv preprint arXiv:2304.10261*, 2023. 1

[4] Y. Dalva, S. F. Altındiş, and A. Dundar, "Vecgan: Image-to-image translation with interpretable latent directions," in *ECCV*. Springer, 2022, pp. 153–169. 1, 3

[5] Y.-J. Chen, S.-I. Cheng, W.-C. Chiu, H.-Y. Tseng, and H.-Y. Lee, "Vector quantized image-to-image translation," in *ECCV*, 2022. 1, 7, 9, 10

[6] L. Zhang, X. Chen, X. Tu, P. Wan, N. Xu, and K. Ma, "Wavelet knowledge distillation: Towards efficient image-to-image translation," in *CVPR*, 2022, pp. 12 464–12 474. 1, 3

[7] M. Ko, E. Cha, S. Suh, H. Lee, J.-J. Han, J. Shin, and B. Han, "Self-supervised dense consistency regularization for image-to-image translation," in *CVPR*, 2022, pp. 18 301–18 310. 1

[8] S. Kim, J. Baek, J. Park, G. Kim, and S. Kim, "Instaformer: Instance-aware image-to-image translation with transformer," in *CVPR*, 2022, pp. 18 321–18 331. 1

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797. 1, 3

[10] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *CVPR*, 2020, pp. 8188–8197. 1, 3, 7, 8, 9, 10

[11] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *NeurIPS*, vol. 30, 2017. 1, 3

[12] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189. 1, 3

[13] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *ICLR*, 2020. 1

[14] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *ICCV*, 2019, pp. 10 551–10 560. 1, 3

[15] K. Baek, Y. Choi, Y. Uh, J. Yoo, and H. Shim, "Rethinking the truly unsupervised image-to-image translation," in *ICCV*, 2021, pp. 14 154–14 163. 1, 3

[16] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. J. Lee, E. Shechtman, and R. Zhang, "Few-shot image generation via cross-domain correspondence," in *CVPR*, 2021, pp. 10 743–10 752. 1, 3

[17] Y. Zhang, Y. Wei, Z. Ji, J. Bai, W. Zuo *et al.*, "Towards diverse and faithful one-shot adaption of generative adversarial networks," in *NeurIPS*, 2022. 1, 3, 4, 7, 9, 10, 14

[18] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *CVPR*, June 2022, pp. 2426–2435. 1, 3, 7, 9, 10

[19] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," in *ICLR*, 2023. 1, 7, 9, 10

[20] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Unsupervised image-to-image translation with generative prior," in *CVPR*, 2022, pp. 18 332–18 341. 2, 3, 7, 9, 10

[21] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *ICLR*, 2019. 2, 8

[22] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *CVPR*, 2020, pp. 2437–2445. 2, 3, 7, 9, 10

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763. 2, 5

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232. 3

[25] C. Jung, G. Kwon, and J. C. Ye, "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks," in *CVPR*, 2022, pp. 18 260–18 269. 3

[26] J. Xiao, L. Li, C. Wang, Z.-J. Zha, and Q. Huang, "Few shot generative model adaption via relaxed spatial structural alignment," in *CVPR*, 2022, pp. 11 204–11 213. 3

[27] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM TOG*, vol. 41, no. 4, pp. 1–13, 2022. 3, 4, 6

[28] P. Zhu, R. Abdal, J. Femiani, and P. Wonka, "Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks," in *ICLR*, 2022. 3, 4, 7

[29] Y. Zhao, H. Ding, H. Huang, and N.-M. Cheung, "A closer look at few-shot image generation," in *CVPR*, 2022, pp. 9140–9150. 3

[30] Y. Zhao, K. Chandrasegaran, M. Abdollahzadeh, and N. man Cheung, "Few-shot image generation via adaptation-aware kernel modulation," in *NeurIPS*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=Z5SE9PiAO4t 3

[31] A. K. Mondal, P. Tiwary, P. Singla, and P. AP, "Few-shot cross-domain image generation via inference-time latent-code learning," in *ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=sCYXJr3QJM8 3

[32] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *CVPR*, June 2021. 4

[33] Y. Xu, Y. Shen, J. Zhu, C. Yang, and B. Zhou, "Generative hierarchical features from synthesizing images," in *CVPR*, 2021. 4

[34] K. Kang, S. Kim, and S. Cho, "Gan inversion for out-of-range images with geometric transformations," in *ICCV*, October 2021, pp. 13 941–13 949. 4, 13

[35] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Restyle: a residual-based stylegan encoder via iterative refinement," in *ICCV*, October 2021. 4

[36] Q. Bai, Y. Xu, J. Zhu, W. Xia, Y. Yang, and Y. Shen, "High-fidelity gan inversion with padding space," in *ECCV*. Springer, 2022, pp. 36–53. 4

[37] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *ICCV*, 2019, pp. 4432–4441. 4

[38] ——, "Image2stylegan++: How to edit the embedded images?" in *CVPR*, 2020, pp. 8296–8305. 4

[39] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM TOG*, vol. 40, no. 3, pp. 1–21, 2021. 4

[40] Y. Xu, Y. Du, W. Xiao, X. Xu, and S. He, "From continuity to editability: Inverting gans with consecutive images," in *ICCV*, 2021, pp. 13 910–13 918. 4

[41] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM TOG*, vol. 40, no. 4, pp. 1–14, 2021. 4

[42] P. Zhu, R. Abdal, Y. Qin, J. Femiani, and P. Wonka, "Improved stylegan embedding: Where are the good latents?" *arXiv preprint arXiv:2012.09036*, 2020. 4, 6

[43] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, "Shifted diffusion for text-to-image generation," in *CVPR*, June 2023, pp. 10 157–10 166. 5

[44] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM TOG*, vol. 42, no. 1, pp. 1–13, 2022. 5, 8

[45] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your gan: Interactive point-based manipulation on the generative image manifold," in *ACM SIGGRAPH*, 2023, pp. 1–11. 5

[46] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *ICCV*, 2021, pp. 2085–2094. 6
[47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. 7, 8
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. 7
[49] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *NeurIPS*, vol. 33, pp. 12 104–12 114, 2020. 8
[50] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410. 8
[51] arfafax, "E621faces," https://github.com/arfafax/E621-Face-Dataset, 2020. 8
[52] Anonymous, D. community, and G. Branwen, "Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset (january 20)," https://gwern.net/Danbooru2020, 2021. 8
[53] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015. 8
[54] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, "Hyperstyle: Stylegan inversion with hypernetworks for real image editing," in *CVPR*, 2022, pp. 18 511–18 521. 8
[55] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012. 8
[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, vol. 30, 2017. 8
[57] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *NeurIPS*, vol. 29, 2016. 8
[58] A. Sauer, K. Schwarz, and A. Geiger, "Stylegan-xl: Scaling stylegan to large diverse datasets," in *ACM SIGGRAPH*, 2022, pp. 1–10. 14
[59] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in *CVPR*, 2023. 14

Yong Du (Member, IEEE) earned the B.Sc. and M.Sc. degrees from Jiangnan University, Wuxi, China, in 2011 and 2014, and the Ph.D. degree from South China University of Technology, Guangzhou, China, in 2019. He is currently an associate professor at the School of Computer Science and Technology, Ocean University of China. His research focuses on computer vision, image processing, machine learning, and generative models. He has published several papers in top venues, including CVPR, ICCV, ECCV, and IEEE Transactions on Image Processing, Multimedia, and Cybernetics.

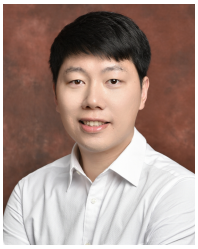
Jiahui Zhan pursued her M.Sc. degree in the School of Computer Science and Technology at Ocean University of China from 2021 to 2024. She will continue her studies toward a Ph.D. degree at Shanghai Jiao Tong University. Her current research interests focus on computer vision and generative models.

Xinzhe Li is an M.Sc. student at the School of Computer Science and Technology, Ocean University of China, Qingdao, China. His research interests include computer vision, image processing, and multi-modal understanding.

Junyu Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, Edinburgh, U.K., in 2003. He is currently a Professor, the Dean of Faculty of Information Science and Engineering, Ocean University of China. His research interests include visual information analysis and understanding, machine learning and underwater image processing.

Sheng Chen (Life Fellow, IEEE) received the Ph.D. degree in control engineering from City University, London, in 1986. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (D.Sc.), from the University of Southampton, Southampton, U.K.. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he holds the post of Professor in intelligent systems and signal processing. His research interests include neural network and machine learning, adaptive signal processing, nonlinear system modeling, and evolutionary computation methods and optimization. He has published over 700 research papers. He is a fellow of the United Kingdom Royal Academy of Engineering and IET.

Ming-Hsuan Yang (Fellow, IEEE) is a professor of Electrical Engineering and Computer Science at the University of California, Merced. Yang serves as a program co-chair of the IEEE International Conference on Computer Vision (ICCV) in 2019, program co-chair of the Asian Conference on Computer Vision (ACCV) in 2014, and general co-chair of ACCV 2016. Yang served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, and is an associate editor of the International Journal of Computer Vision, Image and Vision Computing, and Journal of Artificial Intelligence Research. He received the NSF CAREER award in 2012 and Google Faculty Award in 2009.



Shengfeng He (Senior Member, IEEE) is an associate professor in the School of Computing and Information Systems at Singapore Management University. He was previously on the faculty of the South China University of Technology from 2016 to 2022. He obtained his B.Sc. and M.Sc. degrees from Macau University of Science and Technology in 2009 and 2011, respectively, and a Ph.D. degree from City University of Hong Kong in 2015. His research interests include computer vision and generative models. He has received awards such as the Google Research Awards, the Best Paper Award at PerCom24, and the Lee Kong Chian Fellowship. He is a senior member of IEEE and CCF. He serves as the lead guest editor of IJCV and as an associate editor for IEEE TNNLS, IEEE TCSVT, Visual Intelligence, and Neurocomputing. He also serves on the area chair/senior program committees of ICML, AAAI, IJCAI, and BMVC.