LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data

Jihye Park *,1, Sunwoo Kim *,1, Soohyun Kim *,1, Seokju Cho 1, Jaejun Yoo 2, Youngjung Uh 3, Seungryong Kim † 1

1 Korea University, Seoul, Korea 2 UNIST, Ulsan, Korea 3 Yonsei University, Seoul, Korea

1 {ghp1112, sw-kim, shkim1211, seokju_cho, seungryong_kim}@korea.ac.kr 2 jaejun.yoo@unist.ac.kr 3 yj.uh@yonsei.ac.kr

Abstract

Existing techniques for image-to-image translation commonly have suffered from two critical problems: heavy reliance on per-sample domain annotation and/or inability to handle multiple attributes per image. Recent truly-unsupervised methods adopt clustering approaches to easily provide per-sample one-hot domain labels. However, they cannot account for the real-world setting: one sample may have multiple attributes. In addition, the semantics of the clusters are not easily coupled to human understanding. To overcome these, we present LANGuage-driven Image-to-image Translation model, dubbed LANIT. We leverage easy-to-obtain candidate attributes given in texts for a dataset: the similarity between images and attributes indicates per-sample domain labels. This formulation naturally enables multi-hot labels so that users can specify the target domain with a set of attributes in language. To account for the case that the initial prompts are inaccurate, we also present prompt learning. We further present domain regularization loss that enforces translated images to be mapped to the corresponding domain. Experiments on several standard benchmarks demonstrate that LANIT achieves comparable or superior performance to existing models. The code is available at github.com/KU-CVLAB/LANIT.

1. Introduction

Unpaired image-to-translation frameworks aim to translate an image from one domain to another [21,31,46]. They are typically trained on many images with their respective domain labels. It has been widely studied for a decade in numerous computer vision applications, such as image manipulation [14,62,66], multi-modal translation [3,9,10,21,35], and instance-aware translation [6,23,30].

* Equal contribution
† Corresponding author

Conventional methods [9,10,21] required at least per-sample domain supervision based on the strict assumption that each image should be described by a one-hot domain label, which is frequently violated; e.g., a face can have multiple attributes. In addition, annotating domain label for each image may become considerably labor-intensive and error-prone, and subjectivity may lead to inconsistency across different annotators.

To mitigate such a heavy reliance on per-sample domain annotation, some recent methods formulate the problem in a few-shot setting [35,51] or semi-supervised learning setting [55]. More recently, TUNIT [3] and Kim et al. [29] present a framework for jointly learning domain clustering and translation within a unified model, thus enabling a truly-unsupervised learning of image-to-image translation. Although these methods ease the burden of per-sample domain annotation and show relatively competitive performance, they still inherit the limitation of using one-hot domain labels to train the image translation model. In addition, these truly-unsupervised learning methods only learn domain clusters that are dominant in the dataset, which sometimes do not reflect semantic meanings. Such limitations make users confused in understanding the semantic meaning of each learned domain cluster.

In this paper, to overcome the problems in per-sample-level and unsupervised learning methods, we propose using the dataset-level annotation for the first time as exemplified in Fig. 1. In practice, it could be much easier to list candidate textual domain descriptions that describe images in the dataset. Using the candidate attributes given in texts for a dataset, we specify the target domain given a reference image according to its similarity to the attributes in vision-language embedding [49]. Then, a generator translates a source image to the target domain using the target style code of the reference image. As similarities naturally lead to a multi-hot label, we can aggregate the style codes of multiple domains to represent the style code for a joint domain. In addition, to account for the fact that the initial prompts may
be sometimes inaccurate, we also propose prompt learning and encourage the translated images to have corresponding labels with our proposed domain regularization loss.

Our proposed framework just requires some textual domain descriptions for defining domains and it can ease the burden on the per-sample annotating process. Thanks to the proposed dataset-level annotation, users can explicitly define the semantic meaning of each cluster, limiting its applicability. Unlike this, our framework can select and train the domains that have explicit semantic meaning, which is more applicable.

Figure 1. Levels of supervision. For unpaired image-to-image translation, (a) conventional methods [10, 21, 35, 55] require at least per-sample-level domain supervision, which is often hard to collect. To overcome this, (b) unsupervised learning methods [3, 29] learn image translation model using a dataset itself without any supervision, but it shows limited performance and lacks the semantic understanding of each cluster, limiting its applicability. Unlike them, (c) we present a novel framework that requires a dataset with possible textual domain descriptions (i.e., dataset-level annotation), which achieves comparable or even better performance than previous methods.

Figure 2. Examples of semantic encoding. The existing unsupervised method [3] allows users to translate one of several clusters. However, the learned clusters lack semantic meaning and sometimes some attributes do not appear in any clusters. Unlike this, our framework can select and train the domains that have explicit semantic meaning, which is more applicable.

2. Related Work

Image-to-Image Translation. While early methods for image-to-image translation considered a paired setting [22], most recent state-of-the-arts focus on an unpaired setting [3, 11, 19, 40, 46, 57, 62, 66]. This trend was initiated by CycleGAN [67], of which many variants were proposed, formulating uni-modal models [34, 60, 61] and multi-modal models [9, 10, 21, 31]. Although these methods overcome the inherent limitation of the paired setting, they still require at least per-sample domain annotation, which is often labor intensive or even impossible to achieve in some cases. In addition, this per-sample supervision does not consider that some images may require multi-hot domain labels such as facial attributes, e.g., blond and young.

To mitigate the reliance on per-sample supervision, FUNIT [35] and CoCo-FUNIT [51] proposed a few-shot setting, but still required a few annotations. SEMIT [55] utilizes a pseudo labeling technique. S3GAN [42], Self-conditioned GAN [37], and [4] adopt clustering methods or a classifier. Recently, TUNIT [3] and Kim et al. [29] have proposed the truly-unsupervised frameworks with clustering approach or prototype-based learning, which do not need any per-sample domain supervision. However, their performance was limited in that it is challenging to understand the semantic meaning of each pseudo domain, which in turn limits their applicability. Note that the aforementioned methods also inherit the problem of the one-hot domain labeling assumption.

Vision-Language Model in Image Manipulation. Many large-scale vision-language models [24, 49] have been popularly adopted in computer vision problems. Especially, CLIP [49] has been employed in image synthesis and manipulation with GANs [1, 38, 47, 56]. For instance, Style-CLIP [47] and HairCLIP [56] manipulate the latent space of StyleGAN [27], through the guidance of CLIP, but they heavily rely on pre-trained StyleGAN. CLIP2StyleGAN extends such an approach in an unsupervised manner that eliminates the need for target style descriptions.

More recently, some works [2, 12, 16, 28, 39, 45, 50] combine a diffusion model [18] with CLIP [49] for text-driven image manipulation with high fidelity. However, they show limited performance in realistically translating in the real image domain. Furthermore, they rely on pre-trained dif-
fusion model [18], and they are only available on text-conditioned manipulation. On the other hand, our LANIT is not dependent on pretrained Large StyleGAN [27] and also learns style manifold from given image dataset and a few numbers of prompts that designate explicit semantics to be learned.

Prompt Learning. Since the success of pretrained large-scale language models such as GPT-3 [8] in NLP, various methods for optimizing the prompt in a text format have been proposed [33, 36, 48, 52]. Inspired by these works, some works [13, 20, 25, 32, 59, 64, 65] using CLIP [49] also attempted to optimize the input prompt. For instance, CoCoOp [65] showed that using the learned continuous prompt could surpass the manually-designed discrete prompt-based method on zero-shot image classification. However, they require class supervision. Recently, several works [20, 53, 63] propose unsupervised prompt tuning. In this work, we explore combining image translation with prompt learning in an unsupervised way for the first time.

3. Methodology

3.1. Motivation and Problem Formulation

We leverage a list of dataset-level domain descriptions in the text form and their similarity with the unlabeled images in CLIP [49] embedding space to measure a pseudo domain label of style image. By doing so, we aim to alleviate difficulty in the labeling process and give explicit semantic meaning in the learned domain for the multi-hot label setting.

One naive approach obtaining the pseudo domain label is by calculating the similarity between image and text features with CLIP [49]. However, we observe that this leads to inaccurate labeling and noisy translation since the image-text feature space in CLIP [49] is not perfectly aligned. To mitigate this issue, we introduce two solutions, an adaptive thresholding with the base prompt for selecting confident pseudo labels, and a prompt learning technique with domain regularization loss for boosting confidence in calculating similarity.

3.2. Image Translation with Pseudo Domain Label

Our model, consisting of content encoder $E_C$, style encoder $E_S$, mapping encoder $E_M$, and generator $G$ similar to StarGAN2 [10], aims to learn an image-to-image translation with content image $X$ and style image $Y$ (or between domains $X'$ and $Y'$) to generate a translated image $\hat{Y}$.

In specific, we first extract the content vector and style vector from $X \in \mathbb{R}^{H \times W \times 3}$ and $Y \in \mathbb{R}^{H \times W \times 3}$ with height $H$ and width $W$ such that $c_x = E_C(X) \in \mathbb{R}^{h_c \times w_c \times c}$ with height $h_c$, width $w_c$, and channel $c$, and $s_y = E_S_n(Y) \in \mathbb{R}^{1 \times s}$ for $n$-th domain (or $s = E_M(z)$ from a random latent $z$), respectively. While existing methods, e.g., StarGAN2 [10], select a single vector from $\{s_1^y, ..., s_N^y\}$ of $N$ domains according to ground-truth one-hot domain label, which is inserted to the generator, we aggregate such style vectors $\{s_1^y, ..., s_N^y\}$ with the pseudo domain label $d^y = [d_n^y]_{n=1}^N \in \mathbb{R}^{N \times 1}$ defined as a binary vector, which will be discussed in the following, such that

$$a^y = \frac{1}{M^y} \sum_{n=1}^N s_n^y d_n^y,$$

where $M^y$ is the number of one values in the pseudo domain label $d^y$. With this aggregated style vector $a^y \in \mathbb{R}^{1 \times s}$ and content vector $c_x$, our networks finally generates the output such that $\hat{Y} = G(c_x, a^y)$, containing a style normalization layer. The detailed network architecture is described in the suppl. material.
3.3. Language-Driven Domain Labeling

In this section, we explain how to obtain the pseudo multi-hot domain label \( \mathbf{d}^s \) for an image \( \mathbf{y} \) using pretrained vision-language model. Unlike TUNIT [3] that clusters the images from the dataset using the visual features only, we use \( n \)-th domain’s textual description as prompt \( \mathbf{t}_n \in \mathbb{R}^{1 \times L} \) which is defined as

\[
\mathbf{t}_n = [p_1, p_2, \ldots, p_L, p^\text{domain}_n],
\]

which consists of one of candidate domains \( \{p^\text{domain}_n\}_{n=1}^N \) and a template \( p_l \) \((l \in \{1, \ldots, L\})\) that is shared with all the candidate domains, where \( p_l \) is a vector for a word with the same dimension as \( p^\text{domain}_n \). For instance, all \( p_l \) and \( p^\text{domain}_n \) can be initialized with a template, e.g., “a face with”, and given textual domain descriptions, e.g., “black hair”, respectively. It is utilized to softly cluster the images by measuring the similarity between visual features and language features. In our framework, all the prompts \( \mathbf{T} = \{\mathbf{t}_n\}_{n=1}^N \in \mathbb{R}^{N \times (L+1)} \) are initialized by the human-annotated domain descriptions or keywords, namely dataset-level supervision.

Based on this definition, we first extract vision and language features \( \mathbf{v}^y = \mathbf{E}_V(\mathbf{y}) \in \mathbb{R}^{1 \times k} \) with \( k \) channels, and \( \{\mathbf{u}_n\}_{n=1}^N = \mathbf{E}_L(\mathbf{T}) \in \mathbb{R}^{N \times k} \) with \( k \) channels from style \( \mathbf{y} \) and all the prompts \( \mathbf{T} \), respectively, by using pretrained vision-language model, i.e., CLIP [49]. Note that any vision-language models with vision and language encoders, i.e., ALIGN [24], can be used in our framework. We then measure a similarity \( f_n^x = \mathbf{f}_n^x \in \mathbb{R}^{N \times 1} \), computed as:

\[
f_n^x = \mathbf{v}^y \cdot \mathbf{u}_n^t,
\]

where \( \mathbf{v} = \mathbf{v}/\|\mathbf{v}\| \) and \( \mathbf{u} = \mathbf{u}/\|\mathbf{u}\| \).

By utilizing the measured similarity \( f_n^x \), we can obtain the multi-hot pseudo domain label \( \mathbf{d}^s \) with a simple candidate selecting method, e.g., top-\( K \) operation or thresholding. However, with top-\( K \) operation, we can just select \( K \) number of attributes ignoring the other attributes that would exist in the style image. Moreover, a simple thresholding with a predefined hyperparameter would limit the performance. To overcome these, we present an adaptive thresholding technique based on the observation that the more specified text descriptions, e.g., “a face with blonde hair” make higher similarity with corresponding images than the base prompt as a neutral text description, e.g., “a face with”, that represent each dataset. We set the base prompt \( \mathbf{p} \) as the template, meaning that they are updated simultaneously, and thus \( \mathbf{u}_n^p = \mathbf{E}_L(\mathbf{p}) \in \mathbb{R}^{1 \times k} \). We then define the domain label \( \mathbf{d}^y \in \{\mathbf{d}^y\}_{n=1}^N \in \mathbb{R}^{N \times 1} \) as follows:

\[
d_n^y = \begin{cases} 
1, & \text{if, } f_n^x \geq \mathbf{v}^y \cdot \mathbf{u}_n^p, \\
0, & \text{otherwise.}
\end{cases}
\]

3.4. Prompt Learning

So far, we have discussed the method for language-driven domain labeling and image translation with the pseudo domain label. However, since the pre-defined templates may not be optimal to describe all the images in the dataset, the performance would be noisy and limited. To overcome these, we present a prompt learning technique as shown in Fig. 4.

Similar to [20, 64, 65], we set the prompt \( \mathbf{t}_n \) as learnable except \( p^\text{domain}_n \). In other words, the templates \( \{p_l\}_l \) are set as the learnable parameters in \( \mathbf{t}_n \). By allowing the template to be updated, more optimal prompts can be optimized, which in turn boosts the translation performance. Note that the input domain descriptions should be given in our framework, but they can also be obtained by using a predefined dictionary [1, 49] in a manner that the highly relevant texts to the dataset can be selected based on the similarity between an image and candidate texts.

3.5. Loss Functions

Adversarial Loss. We adopt an adversarial loss to guide the translated images \( \hat{\mathbf{y}} \) to arrive at the training distribution of the target domain. Following StarGAN2 [10] and TUNIT [3], we also adopt the multi-domain discriminators, but we use the outputs of multi-domain discriminators weighted
by a multi-hot domain label $d^y$:

$$L_{adv} = \mathbb{E}_{x, y} \sum_{n=1}^{N} \log D_n(y)d^n_x + \log(1 - D_n(G(x, a^y))d^n_x),$$

(5)

where $D_n(\cdot)$ denotes $n$-th discriminator output and $d^n_x$ denotes $n$-th element of $d^y$. It should be noted that if we set $d^y$ as a one-hot domain label by ground-truth or pseudo-label, this loss function becomes the same adversarial loss in [3, 10].

**Domain Regularization Loss.** We further present a domain regularization loss that encourages the networks to generate an image $\hat{y}$ following the domain label of input style image $y$ more precisely. To this end, perhaps one of the simplest functions is to use the dot product between $d^y$ and $f^y$ such that $\mathcal{L} = d^y \cdot f^y$. Minimizing this, however, can induce erroneous solutions, since considering an inactive domain label, i.e., $d^n_x = 0$, may degrade the performance further. To overcome this, we present a novel loss function. We first make a domain label pair, $d^y$ and $d_{inv}^y(n)$, which has same labels with $d^n_y$ except the $n$-th label by replacing $n$-th domain label $d^n_x$ to $d_{inv}^y(n) = 1 - d^n_x$. We then generate $\hat{y}' = G(e^x, a_{inv}^y)$ where $a_{inv}^y$ is obtained using $\hat{y}$ and $d_{inv}^y(n)$. We then encourage not only $f^y$ to follow $d^y$ at $n$-th label, but also $f_{inv}^y$ to follow $d_{inv}^y(n)$ at $n$-th label, such that

$$L_{dl} = \mathcal{H}(d^n_x, f^n_y) + \mathcal{H}(d_{inv}^y(n), f_{inv}^y(n)), \quad (6)$$

where $\mathcal{H}(\cdot, \cdot)$ denotes a binary cross-entropy.

**Cycle-Consistency Loss.** To make the translated image similar to its original image, which also regularizes the ill-posed translation problem, we adopt a cycle-consistency loss such that

$$L_{cyc} = \mathbb{E}_{x, y} ||x - G(\hat{y}, a^x)||_1, \quad (7)$$

where $e^x$ denotes the content from $\hat{y}$, and $a^x$ denotes the style vector of input $x$. By encouraging the generator $G$ to reconstruct the input image $x$ with estimated style vector $a^x$, $G$ learns to preserve the original characteristics of $x$.

**Style Reconstruction Loss.** While the generator is able to synthesize realistic images with the losses above, the synthesized results are not guaranteed to be style-consistent with $y$. In order to better learn style representation, we compute $l_1$ loss between the style vector from the translated image and style image such that

$$L_{sty} = \mathbb{E}_{x, y} ||s^x - \mathcal{S}(\hat{y})||_1. \quad (8)$$

**Style Diversification Loss.** To learn the generator to produce diverse images, we employ a diversity sensitive loss [43, 58] defined as:

$$L_{ds} = \mathbb{E}_{x, y} \max(|\mathcal{G}(x, \mathcal{E}_M(z_1)) - \mathcal{G}(x, \mathcal{E}_M(z_2))|), \quad (9)$$

where $z_1$ and $z_2$ are random latent vectors from Gaussian distribution. Note that we only adopt the diversification loss using random vectors, not the style vector $a^y$ from an image. Also, we utilize the domain label $d^y$ as the style image $y$ as a pseudo-label for the latent vector.

**Overall Objective.** Full loss functions are as follows:

$$L_{total} = \lambda_{adv}L_{adv} + \lambda_{dl}L_{dl} + \lambda_{cyc}L_{cyc} + \lambda_{sty}L_{sty} - \lambda_{ds}L_{ds}, \quad (10)$$

where $\lambda_{adv}, \lambda_{dl}, \lambda_{cyc}, \lambda_{sty}$, and $\lambda_{ds}$ are hyper-parameters.

### 4. Experiments

**Implementation Details** We adopted StarGAN2 [10] as the baseline architecture for content, style, mapping encoders, generator, and discriminator. Additional implementation details will be explained in the supplementary material.

**Setting up Templates and Candidate Domains.** We set the candidate domains by randomly sampling among the domain or attribute names in each dataset. For a fair comparison, we experiment three times with different domains respectively to measure quantitative results. The selected candidate domains are shown in the supplementary materials. We refer to the templates from [5, 12] and slightly modify them to align for each dataset. To improve the accuracy of pseudo labels, we also adopt template augmentation in our baseline model.

**Datasets.** We conduct the experiments on four standard datasets, including Animal Faces-10, Food-10, CelebA-HQ, and LHQ [7, 35, 41, 54]. In particular, we consider 10 textual domain descriptions among Animal Faces-149 and Food-101, respectively, selected in TUNIT [3], to define the initial prompt of our framework. For CelebA-HQ, we obtain 40 attributes, and sample 10 attributes for three times to define the initial prompt and report the average results. Since LHQ dataset [54] does not provide any attributes, we sampled various target domain descriptions and selected 10 domains for experiments.

**Evaluation Metrics.** In our experiments, we adopt three quantitative evaluation metrics. First, the mean of class-wise Fréchet Inception Distance (mFID) [17] is used to evaluate how the translated image reflects the input style. In addition, we use Density and Coverage (D&C) [44], which measure the fidelity and diversity of the translated images, respectively. The lower the mFID value, the better quality of the image. As fidelity and diversity get better, the D&C scores are higher or closer to 1. We also measure Acc., namely classification accuracy, between the domain labels with the highest probability and ground-truth domain labels.
### 4.1. Comparisons for Other Models

**Quantitative Results.** We first report the quantitative comparison of our model with StarGAN2 [10], Smoothing [40] (supervised), TUNIT [3] and Kim *et al.* [29] (unsupervised) on CelebA-HQ, AnimalFaces-10 and Food-10 datasets. Note that StarGAN2 and Smoothing are trained using 10 classes of ground-truth per-sample domain labels. Our and truly-unsupervised models use GT labels for calculating quantitative metrics. Tab. 1 shows the quantitative comparison in terms of mFID, D&C. metrics. LANIT shows the best mFID and D&C scores, proving that our model outperforms the state-of-the-art methods in terms of image quality, fidelity and diversity. We also observe that LANIT consistently outperforms unsupervised models by a large margin on all three datasets on Acc. metric. Although our LANIT only utilizes unlabeled datasets with dataset-level text supervision, it shows competitive results against StarGAN2 [10] and Smoothing [40], which require ground-truth per-sample domain labels. Note that we do not provide quantitative results on LHQ since it does not provide ground-truth labels.

**Qualitative Results.** We show visual results for Tab. 1 in Fig. 5. We observe that the quality differences between our LANIT and TUNIT [3] or Kim *et al.* [29] are significant. The images generated by the other methods tend to contain artifacts and show limited performance in reflecting diverse attributes from style images, while our method yields stable performance in terms of style representation and semantic consistency. Especially, we show that our model is highly capable of capturing multiple subtle attributes, such as “smiling” or “straight hair”, while TUNIT [3] and Kim *et al.* [29] often fail. Interestingly, our LANIT shows competitive results compared to StarGAN2 [10] and Smoothing [40] which are supervised methods, even with better

---

**Table 1. Quantitative comparison on CelebA-HQ [41], AnimalFaces-10 [35] and Food-10 [7]**

The configurations of StarGAN2 and Smoothing use ground-truth domain labels while TUNIT and Kim *et al.* use pseudo-labels generated from each image. Our LANIT uses only textual domain descriptions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mFID ↓</td>
<td>D&amp;C ↑</td>
<td>mFID ↓</td>
</tr>
<tr>
<td>StarGAN2 [10] (sup.)</td>
<td>32.16</td>
<td>1.22 / 0.44</td>
<td><strong>33.67</strong></td>
</tr>
<tr>
<td>Smoothing [40] (sup.)</td>
<td>35.93</td>
<td><strong>1.25 / 0.43</strong></td>
<td>38.93</td>
</tr>
<tr>
<td>TUNIT [3] (unsup.)</td>
<td>61.29</td>
<td>0.24 / 0.13</td>
<td>47.70</td>
</tr>
<tr>
<td>Kim <em>et al.</em> [29] (unsup.)</td>
<td>41.33</td>
<td>0.60 / 0.24</td>
<td>36.83</td>
</tr>
<tr>
<td>LANIT</td>
<td>27.96</td>
<td>0.91 / 0.34</td>
<td>34.11</td>
</tr>
</tbody>
</table>
style preservation. We also show domain-to-domain diverse image synthesis results in Fig. 6, which is guided by target domain labels. Thanks to our proposed prompt learning, the mapping network and style encoder can faithfully produce the style vectors reflecting the target multiple styles.  

**Comparisons to other CLIP-based methods.** We compare our LANIT with CLIP-based text-guided manipulation models, DiffusionCLIP [28] and StyleCLIP [47], on CelebA-HQ in Fig. 7. In the results, DiffusionCLIP and StyleCLIP show a limited capacity to synthesize realistic results with multiple attributes, while our LANIT generates high-quality images, faithfully representing each domain description, regardless of the number of domain descriptions. The reason is that our model focuses on learning style manifold during training from each target domain defined by textual descriptions, while others are designed to just fine-tune the pre-trained generative models on each instance independently. We observe that LANIT also works on a flexible number of K during inference time, which demonstrates the controllability of our model.

**4.2. Ablation Study and Analysis**

**Number of Domain Descriptions N.** In Tab. 2, We validate the effects of the number of candidate domains N compared to TUNIT [3]. Our results show that a different number of N brings minor changes to our model than TUNIT [3], which justifies the robustness of our framework. In animal faces-10, LANIT shows the best performance when there exist 10 candidate domains, whereas in CelebA-HQ, there is little difference in the FID score when N is over 10.

**Predefined Dictionary.** We also evaluate the quantitative results when using a predefined dictionary instead of exploiting specific domain descriptions that we manually provide. For this experiment, we regard the whole class names in each dataset as a dictionary. For sampling candidate domains, we first calculated similarity values on the whole images and target domains, and then only averaged the similarity values assigned to the target domains in the dictionary. Then, we ordered average values in descending order and selected 10 domains. While our LANIT shows a slight drop in both mFID and F1 scores as shown in the Tab. 3 for such a setting in all the datasets, it still records competitive performance compared to the default setting.

**Adaptive Thresholding with Base Prompt.** We study the effectiveness of the adaptive thresholding technique using base prompt compared to the simple top-K method in Tab. 4. We find that the proposed technique brings a large improvement in performance. This indicates that it enables the model to mitigate the influence of ambiguous or noisy

<table>
<thead>
<tr>
<th>Content</th>
<th>“blond hair”</th>
<th>“blond hair”</th>
<th>“bang”</th>
<th>“smiling”</th>
<th>“lipstick”</th>
</tr>
</thead>
<tbody>
<tr>
<td>“mountain”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“waterfall”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“summer”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“ocean”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“cloudy”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7. Additional qualitative results by (from top to bottom) DiffusionCLIP [28], StyleCLIP [47], and our LANIT.

<table>
<thead>
<tr>
<th>Table 3. Quantitative results with pre-defined dictionary.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datasets</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>mFID ↓</td>
</tr>
<tr>
<td>CelebA-HQ [15]</td>
</tr>
<tr>
<td>AnimalFaces-10 [35]</td>
</tr>
<tr>
<td>Food-10 [7]</td>
</tr>
</tbody>
</table>

Figure 6. Latent-guided diverse image synthesis results by our LANIT on LHQ [54]. Target domain text descriptions are shown above each image.

Table 2. Quantitative comparison of LANIT with TUNIT [3] by varying the number of domains.

<table>
<thead>
<tr>
<th>N</th>
<th>Method</th>
<th>AnimalFaces-10 [35]</th>
<th>CelebA-HQ [41]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mFID ↓</td>
<td>D&amp;C ↑</td>
<td>mFID ↓</td>
</tr>
<tr>
<td>4</td>
<td>TUNIT</td>
<td>77.7 / 0.74</td>
<td>61.5</td>
</tr>
<tr>
<td></td>
<td>LANIT</td>
<td>71.6 / 0.46</td>
<td>49.3</td>
</tr>
<tr>
<td>7</td>
<td>TUNIT</td>
<td>62.7 / 0.73</td>
<td>54.7</td>
</tr>
<tr>
<td></td>
<td>LANIT</td>
<td>49.9 / 0.66</td>
<td>43.2</td>
</tr>
<tr>
<td>10</td>
<td>TUNIT</td>
<td>47.7 / 0.81</td>
<td>61.3</td>
</tr>
<tr>
<td></td>
<td>LANIT</td>
<td>34.1 / 0.89</td>
<td>27.9</td>
</tr>
<tr>
<td>13</td>
<td>TUNIT</td>
<td>56.8 / 0.72</td>
<td>98.9</td>
</tr>
<tr>
<td></td>
<td>LANIT</td>
<td>30.1 / 0.85</td>
<td>34.8</td>
</tr>
<tr>
<td>16</td>
<td>TUNIT</td>
<td>54.1 / 0.78</td>
<td>127.7</td>
</tr>
<tr>
<td></td>
<td>LANIT</td>
<td>35.8 / 0.82</td>
<td>27.9</td>
</tr>
</tbody>
</table>

Table 3. Quantitative results with pre-defined dictionary.
data. Additionally, we evaluate our model with single or multiple activated top-K attributes. We can see that the result of Top-3 significantly surpasses Top-1. This demonstrates that an image can be described more precisely when given with multiple keywords or text descriptions, while most prior works including TUNIT, Kim et al. and StarGAN2 assume to assign a single label for each image. In addition, we evaluate the qualitative results in Fig. 8 corresponding to the Tab. 4. It shows that our LANIT captures more diverse attributes rather than Top-1 and Top-3 models.

**Domain Regularization Loss.** As shown in Tab. 4 (D and E), our LANIT improves performance Domain Regularization (DL) loss compared to the baseline on CelebA-HQ. Especially, D&C metrics are significantly increased by adding our proposed domain regularization loss. In Fig. 8, the generated image has the style of reference image more faithfully by using domain regularization loss.

**Prompt Learning.** In order to evaluate the effectiveness of prompt learning, we compare the estimated pseudo-domain labels with and without the proposed prompt learning, where the latter can be regarded as an offline clustering. It shows that with prompt learning, the probability of an image/text prompt pair is more accurately estimated. Since prompt learning helps reduce label noise by refining the prompt, the result without prompt learning shows limited clustering performance, which is verified in Tab. 4 (E). In addition, we further analyze the effect of when to start prompt learning Tab. 5. The stage-wise learning, where we start prompt learning after convergence of the generative model, shows better performance than end-to-end learning with prompt learning. The reason is that the well-generated image corresponding to a given prompt has a critical role in prompt learning. Note that the base prompt is also updated with prompt learning simultaneously.

### 5. Conclusion

In this paper, we propose LANGUAGE-driven Image-to-image Translation framework for Unlabeled Data (LANIT) that leverages dataset-level supervision. By using textual domain descriptions as dataset-level supervision, our model not only mitigates the heavy dependency on per-sample supervision but also considers multi-hot domain labels for an image. The proposed labeling method allows users to explicitly define the semantic meaning of the target domain, which was not possible in the previous truly-unsupervised I2I framework. In addition, as we observe that a naive adoption of a pre-trained vision-language model has limited controllability and even drops the performance of the model, we present some techniques to overcome such limitations: adaptive thresholding, domain regularization loss, and prompt learning. Our experiments show that our LANIT performs better than truly-unsupervised methods in controllability on the application. Furthermore, the quantitative results of our framework have scored competitive or even better values than the existing image-to-image translation models that require per-sample supervision.

**Acknowledgements.** This research was supported by the MSIT, Korea (IITP-2023-2020-0-01819, ICT Creative Consilience program, No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques), and National Research Foundation of Korea (NRF2021R1C1C1006897).
References


