2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC) | 979-8-3315-3047-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/MCSoC64144.2024.00022

VLC-UNIT: Unsupervised Image-to-Image Translation with Vision-Language Classification

Yuying Liang
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
yuyingliang@e.gzhu.edu.cn

Huakun Huang
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
huanghuakun@gzhu.edu.en

Bin Wang
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
dvl.wangbin@e.gzhu.edu.cn

Lingjun Zhao
Electronics and Information
Guangdong Polytechnic Normal University
Guangzhou, China
zhaolj@gpnu.edu.cn

Jiantao Xu
Graduate School of Computer Science and Engineering
University of Aizu
Japan
d8252108@u-aizu.ac.jp

Chen Zhang
Graduate School of Computer Science and Engineering
University of Aizu
Japan
d8252109@u-aizu.ac.jp

***Abstract*—In recent years, unsupervised image-to-image translation technology has garnered widespread attention for its creative expression, accompanied by enhanced data augmentation capabilities. However, current methods are heavily influenced by dataset distributions, resulting in reduced performance on data-scarce samples and inconsistencies in fine-grained categories and target styles, leading to the generation of pseudo categories. To tackle these challenges, we introduce a novel framework called Vision-Language Classification (VLC-UNIT) for UNsupervised Image-to-Image Translation. By leveraging prompts from large vision-language models, VLC-UNIT enhances the semantic understanding of images and improves the consistency and accuracy of image translations. Experimental results demonstrate our VLC-UNIT outperforms existing state-of-the-art techniques.**

***Index Terms*—GAN, Unsupervised Image-to-Image Translation, Vision-language model**

I. Introduction

With the advancement of large-scale models, image translation technology has gained prominence for its remarkable artistic output and has become increasingly pivotal in data augmentation [1]. Within this field, unsupervised image-to-image translation (UNIT) stands out for its capability to operate without requiring manual labeling or data pairing, marking a significant research direction [2] [3] [4]. UNIT aims to preserve the core content of an image while altering its stylistic attributes, facilitating transformations across different domains. Here, 'content' includes orientation, posture, facial features, and approximate shape correspondence, while 'style' refers to visual appearance and artistic expression. This technology generates multiple versions of a single photograph, encompassing different orientations, expressions, and artistic styles, facilitating data augmentation applications across various fields.

Numerous studies can produce high-fidelity images but encounter difficulties with large cross-domain translations. To tackle this, BigGAN [5] employed large-scale GANs for processing but often struggled with quality and intra-domain diversity. GP-UNIT [6] leveraging pre-trained content encoders to facilitate large cross-domain image translation. However, the generated results heavily hinge on dataset quality. As shown in Fig. 1 insufficient samples introduced style-inconsistencies results, diminishing realism and diversity in long-tail datasets, thereby limiting the applicability of image translation technologies. To address these challenges, PFC-UNIT [7] advocates for multiple pre-trained fine-grained classifiers to steer translation. However, adapting to new tasks within the same domain incurs additional costs for pre-training. Furthermore, these pre-trained fine-grained classifiers rely on predefined labels as supervision signals, limiting their applicability beyond the initial pre-training dataset. For example, a fine-grained classifier pre-trained on a dataset of cats excels in guiding the translation of those species but may encounter difficulties when applied to other species. This constraint curtails the effectiveness of such fine-grained models in broader

This work was supported in part by the Natural Science Foundation of Guangdong Province of China (Grant No. 2022A1515010163), in part by the GuangDong Basic and Applied Basic Research Foundation (Project No. 2021A1515110455).

2771-3075/24/$31.00 ©2024 IEEE
DOI 10.1109/MCSoC64144.2024.00022

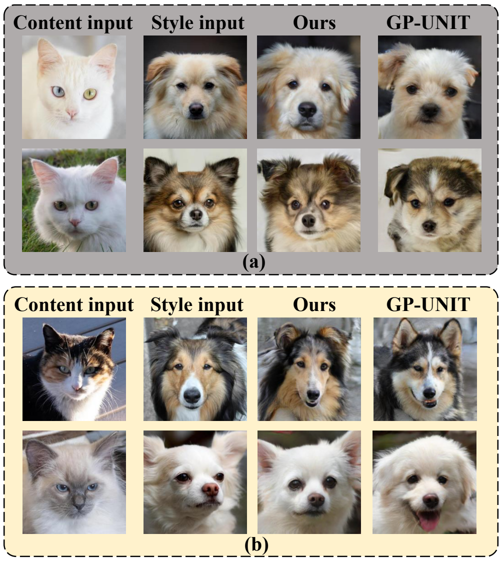

Fig. 1. Limitations of GP-UNIT: (a) Style information loss: fails to generate realistic outcomes when translating the insufficient samples. (b) Style inconsistencies: fails to generate dog breeds that accurately match the nuanced stylistic input.

UNIT applications. In light of these limitations, the rise of the Contrastive Language-Image Pre-training (CLIP) [8], has revolutionized the approach to various downstream tasks due to its powerful capabilities. The impressive performance of CLIP in these tasks has provided us with a new perspective for addressing fine-grained issues in UNIT.

To address the aforementioned challenges, we propose an unsupervised image-to-image translation method with vision-language classification (VLC-UNIT). By feeding multiple fine-grain detailed prompts into the text encoder of the CLIP model, we utilize the extracted text embeddings to guide the image translation process. This method aims to precisely align the translated outputs with the stylistic characteristics of the style target. With textual guidance, the generated outputs tend towards distinct fine-grained categories, enhancing both the realism and stylistic diversity of the results beyond reliance on the input dataset alone. Leveraging CLIP's zero-shot learning capability enables VLC-UNIT to handle new categories. For example, even without direct exposure to zebras, VLC-UNIT can infer their specific characteristics based on descriptions such as horse-like shapes, tiger-like stripes, and panda-like colors during the transformation process. By leveraging CLIP's ability to understand and associate visual and textual information through its extensive pre-training, we can achieve more refined control in UNIT tasks. Experimental results demonstrate that VLC-UNIT surpasses state-of-the-art methods.

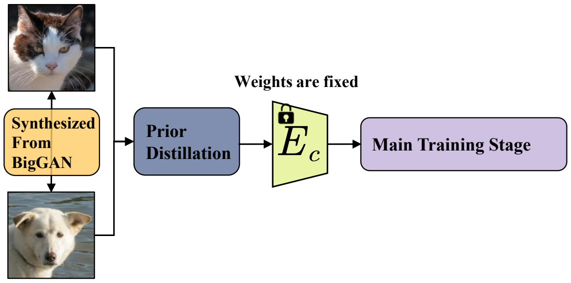

Fig. 2. The training process begins with pre-training for prior distillation, following the GP-UNIT approach, to obtain E_c We then fix E_c during the main training stage.

II. Related work

A. Unsupervised image-to-image translation

In recent years, various methods have been proposed to address UNIT tasks across different domains. Among them, TraVeLGAN [9] introduced Siamese networks to encode images into latent vectors for translation across various domains. U-GAT-IT [10] utilizes attention modules with auxiliary classifiers to guide models to focus on crucial areas between source and target domains. MUNIT [2] decomposes images into content and style spaces to facilitate domain adaptation from source to target images. COCO-FUNIT [11] computes style embeddings conditioned on input images to generalize models to unseen domains. StarGAN2 [12] utilizes style encoding for diverse image generation. However, these methods suffer from significant degradation in performance with large domain translation tasks. BigGAN [5] was introduced using big batch size to handle large cross-domain transformation tasks. GP-UNIT [6] utilized pre-trained content encoders to guide image translation processes but encountered inconsistencies with target styles. However, these methods suffer from style inconsistency, leading to unrealistic representations. PFC-UNIT [7] employed pre-trained fine-grained classifiers to facilitate larger cross-domain translation, but requires additional training for each specific task and remains limited by dataset quality.

B. Vision-language modle

VirTex [13] employs semantically dense captions for training visual representations with a self-regressive prediction method on COCO Captions. ICMLM [14] utilizes a cloze-style approach to pre-training, learning visual representations through caption annotations. ConVIRT [15] employs contrastive learning to derive medical visual representations from naturally paired image-text data without additional expert input. CLIP [8] utilizes contrastive learning to align images and text within a shared vector space, trained on 4 billion image-text pairs. Through the use of textual prompts, CLIP enhances semantic understanding between images and demonstrates impressive zero-shot capabilities.

III. Proposed method

In large cross-domain translation tasks, integrating content and style presents a significant challenge. Inspired by GP-

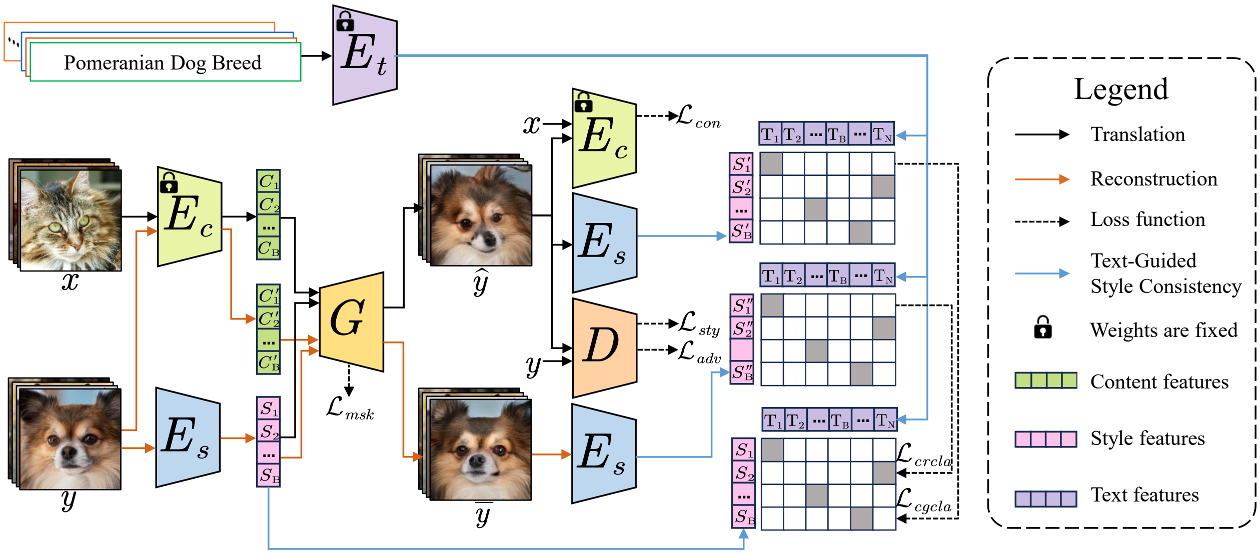

Fig. 3. In the main training stage, the proposed method integrates a pre-trained text encoder with contrastive learning to align the style features of translated images with fine-grained textual categories, ensuring style consistency and accurate representation across diverse classifications such as animal breeds, vehicle types, and facial attributes.

UNIT [6], we implement prior distillation before the main training phase by feeding images synthesized from the identical latent code of BigGAN, as depicted in Fig. 2. This preliminary step primes the content encoder E_c, enabling extraction of abstract cross-domain content features, which are then fixed during the main training stage to preserve its efficacy.

As shown in Fig. 3, during the main training phase, an image x serves as the content input and an image y as the style input.x is processed by the pre-trained E_C (with fixed weights) to yield content features $C_1, C_2, \ldots, C_B$. Meanwhile, y is input into the style encoder E_s to derive style features $S_1, S_2, \ldots, S_B$. These features are then fed into the generator G to produce the translation result $\hat{y}$. The $\hat{y} = G([C_1, C_2, \ldots, C_B], [S_1, S_2, \ldots, S_B])$ aims to preserve the content of x while conforming to the style of y, guided respectively by a content loss $\mathcal{L}con$ and a style loss $\mathcal{L}style$:

$$\mathcal{L}_{con} = \mathbb{E}_{x,y}\left[\, \|E_c(\hat{y}) - E_c(x)\|_1 \,\right], \tag{1}$$

$$\mathcal{L}_{sty} = \lambda_t \mathbb{E}_{x,y}\left[\, \|f_D(\hat{y}) - f_D(y)\|_1 \,\right], \tag{2}$$

where λ_t is a style weight factor. The discriminator D evaluates $\hat{y}$ to distinguish between real and generated images using the traditional GAN loss $\mathcal{L}_{adv}$, as:

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

TABLE I
FINE-GRAINED TEXT DESCRIPTIONS

Task	Fine-grained Categories
Dog	German Shepherd Dog Breed, Poodle Dog Breed, Pomeranian Dog Breed, Labrador Retriever Dog Breed...
Cat	American Short Hair Cat Breed, Persian Cat Breed, Maine Coon Cat Breed, British Short Hair Cat Breed...
Bird	American Robin Bird Breed, Barn Swallow Bird Breed, House Sparrow Bird Breed, Northern Cardinal Bird Breed...
Car	Sedan, Hatchback, Sport Utility Vehicle, Microcar, Crossover, Pickup Truck, Sports Car, Convertible, Van...
Human Face	Child, Teenager, Young adult, Middle-Aged People, Late Adulthood People, Elderly People, Baldhead People
	Angry Face, Disgust Face, Fear Face, Happy Face, Neutral Face, Sad Face, Surprise Face...

A. Text-Guided Style Consistency

We propose feeding $\hat{y}$ into E_S to obtain the style features $S'_1, S'_2, \ldots, S'_B$ of the translated images. We conducted research and compiled the textual names of fine-grained breeds for dogs, cats, and birds, totaling 120, 68, and 602 breeds respectively. Regarding vehicles, we identified 36 types based on body type and specialized functions. For facial classification, we utilize dual prompts categorizing age group and facial expression, consisting of 10 and 7 categories respectively. Multiple above text prompts representing fine-grained categories are fed into the text encoder E_t to derive the text features $T_1, T_2, \ldots, T_B, \ldots, T_N$, where N denotes the total number of fine-grained categories. Notably, the pre-trained text encoder from the CLIP model is utilized here. We propose to employ contrastive learning to align image features closely with text features corresponding to the target style's fine-grained category, guiding the style features of the translated image towards the target style features. Specifically, this involves computing similarity scores between different image features

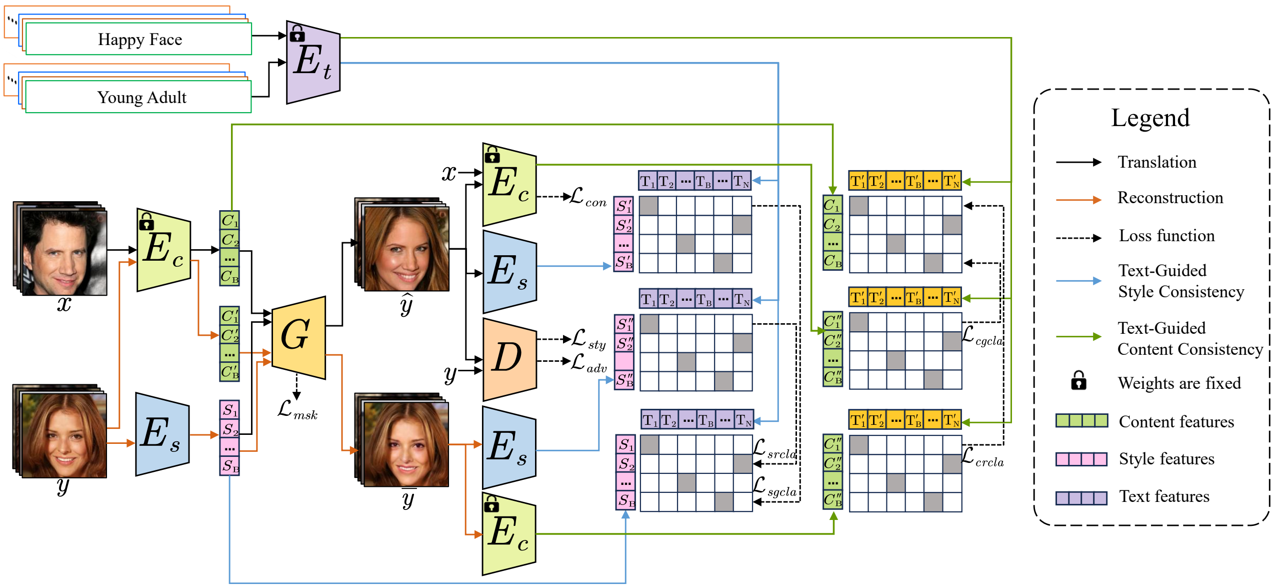

Fig. 4. In the facial transformation tasks between men and women, the proposed method not only ensures style consistency but also maintains age and content characteristics, such as expressions and wrinkles, through the use of age and expression-related text prompts, guiding the transformation process to preserve detailed content information.

$S_1, S_2, \ldots, S_B$, and text features $T_1, T_2, \ldots, T_B, \ldots, T_N$, as:

$$Sim_{y_s} = \mathbf{S} \cdot \mathbf{T}^\mathsf{T} = \sum_{i=1}^{n} S_i T_i^\top. \tag{4}$$

The larger Sim_y indicates a stronger correspondence between the image and the text, whereas a smaller Sim_y implies a weaker correspondence between them. Similarly, we compute similarity scores between image features $S'_1, S'_2, \ldots, S'_B$ and text features $T_1, T_2, \ldots, T_B, \ldots, T_N$ using matrix multiplication, as:

$$Sim_{\hat{y}_s} = \mathbf{S}' \cdot \mathbf{T}^\mathsf{T} = \sum_{i=1}^{n} S'_i T_i^\top. \tag{5}$$

Subsequently, we propose utilizing the generation classification loss λ_{gcla}, employing cross-entropy to measure the discrepancy between the actual and predicted probability distributions, as depicted by:

$$\mathcal{L}_{sgcla} = -\lambda_f \sum_{i=1}^{B} Sim_{y_s}(i) \log Sim_{\hat{y}_s}(i), \tag{6}$$

where λ_f is a weight factor. Smaller $\mathcal{L}_{gcla}$ indicates closer alignment between the fine-grained categories of the translation results $\hat{y}$ and those of the target style y. Furthermore, y undergoes processing by E_c to extract the content features $C'_1, C'_2, \ldots, C'_B$ specific to the style input. The reconstructed output $\bar{y} = G([C'_1, C'_2, \ldots, C'_B], [S_1, S_2, \ldots, S_B])$ aims to faithfully replicate the content and style of y. To obtain the style features $S''1, S''2, \ldots, S''B$ of the recreated images, we propose feeding $\bar{y}$ into E_S. Similarity scores between distinct image features $S''1, S''2, \ldots, S''B$ and text features $T_1, T_2, \ldots, T_B, \ldots$ are computed as follows:

$$Sim_{\bar{y}_c} = \mathbf{S}'' \cdot \mathbf{T}^\mathsf{T} = \sum_{i=1}^{n} S''_i T_i^\top. \tag{7}$$

These scores are used to compute cross-entropy losses λ_{rcla}, which is shown as:

$$\mathcal{L}_{srcla} = -\lambda_q \sum_{i=1}^{B} Sim_{y_s}(i) \log Sim_{\bar{y}_s}(i), \tag{8}$$

where λ_q is a weight factor. As the PFC-UNIT, we use the $\mathcal{L}_{msk}$ for dynamic skip connections in later stages to enhance detailed translation. The objective function of the main training stage is defined as:

$$\begin{aligned} \min_{G,E_s} \max_{D} \mathcal{L}_{adv} + \mathcal{L}_{sty} + \mathcal{L}_{con} + \mathcal{L}_{msk} \\ + \mathcal{L}_{rec} + \mathcal{L}_{srcla} + \mathcal{L}_{sgcla}. \end{aligned} \tag{9}$$

B. Text-Guided Content Consistency

In contrast to other translation tasks, in Male ↔ Female facial transformation tasks, the focus extends beyond style consistency to include matching the age characteristics of the target style. These facial transformation tasks place a greater emphasis on preserving content details, ensuring that the transformed result retains expressions and facial wrinkle patterns. To address these requirements, we modified the network framework, as illustrated in the Fig. 4, demonstrating that text can effectively guide content consistency.

As with the aforementioned process, to achieve style consistency in the transformation results, we employ age-related text prompts to guide the transformation results

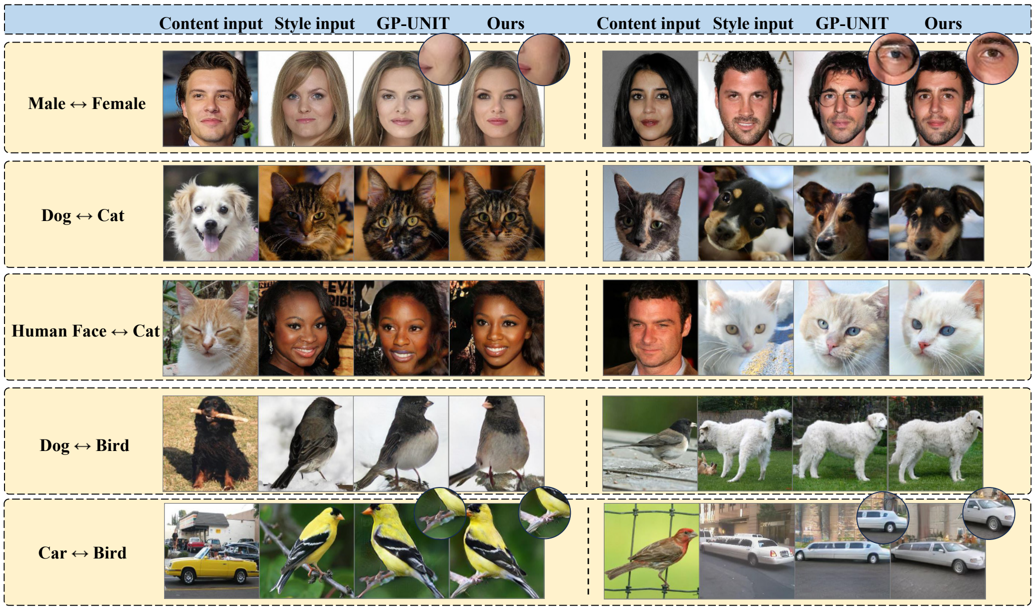

Fig. 5. Visual comparisons of UNIT results across various tasks including Male ↔ Female, Dog ↔ Cat, Human Face ↔ Cat, Dog ↔ Bird, and Car ↔ Bird. The figure represents the content input, style input, results from GP-UNIT, and results from our proposed model. Magnifiers are used on some images to highlight detailed comparisons.

to correspond to the target style's age characteristics. Additionally, we propose using text to guide the retention of content information during transformation. We input expression-related descriptions into the text encoder to obtain text features $T_1', T_2', \dots, T_B', \dots, T_N'$. The generated images are then input into E_c to obtain the translated content features $C_1'', C_2'', \dots, C_B''$. Similarity scores are calculated between $C_1'', C_2'', \dots, C_B''$ and $T_1', T_2', \dots, T_B', \dots, T_N'$, as well as between input content images $C_1, C_2, \dots, C_B$ and $T_1', T_2', \dots, T_B', \dots, T_N'$:

$$Sim_{y_c} = \mathbf{C} \cdot \mathbf{T'}^{\top} = \sum_{i=1}^{n} C_i {T_i'}^{\top}, \tag{10}$$

$$Sim_{\hat{y}_c} = \mathbf{C''} \cdot \mathbf{T'}^{\top} = \sum_{i=1}^{n} {C''}_i {T_i'}^{\top}. \tag{11}$$

After that, we utlize cross-entropy to measure the discrepancy between the actual and predicted probability distributions, as:

$$\mathcal{L}_{cgcla} = -\lambda_f \sum_{i=1}^{B} Sim_{y_c}(i) \log Sim_{\hat{y}_c}(i), \tag{12}$$

where λ_f is a weight factor. The recreation images are inputted into E_c to obtain $C_1''', C_2''', \dots, C_B'''$. Similarly, similarity scores are calculated between $C_1''', C_2''', \dots, C_B'''$ and $T_1', T_2', \dots, T_B', \dots, T_N'$. Cross-entropy is then used to distinguish between the actual content and recreation content distributions, as follows:

$$Sim_{\bar{y}_c} = \mathbf{C'''} \cdot \mathbf{T'}^{\top} = \sum_{i=1}^{n} {C'''}_i {T_i'}^{\top}, \tag{13}$$

$$\mathcal{L}_{crcla} = -\lambda_q \sum_{i=1}^{B} Sim_{y_c}(i) \log Sim_{\bar{y}_c}(i), \tag{14}$$

where λ_q is a weight factor. The objective function of the Male ↔ Female tasks is defined as:

$$\min_{G,E_s} \max_{D} \mathcal{L}_{adv} + \mathcal{L}_{sty} + \mathcal{L}_{con} + \mathcal{L}_{msk} + \mathcal{L}_{srcla} \tag{15}$$
$$+ \mathcal{L}_{sgcla} + \mathcal{L}_{rec} + \mathcal{L}_{crcla} + \mathcal{L}_{cgcla}.$$

This approach ensures that the transformed images not only match the target style in terms of age but also preserve the essential content features, resulting in more accurate and realistic transformations.

IV. Experiments

A. *Experimental Settings*

Datasets. During the pre-training stage, we employed ImageNet291 [6], CelebA-HQ [17], and synImageNet [6], generated from BigGAN [5], to train the coarse content encoder. In the main training stage, we conducted extensive experiments across three datasets for various translation tasks. We utilized the AFHQ [12] dataset for Dog ↔ Cat translations, and the CelebA-HQ [17] dataset for Male $leftrightarrow$ Female translations. ImageNet291 [6] was used for Bird ↔ Car translations and Bird ↔ Dog translations, leveraging the respective categories. Additionally, the AFHQ cat categories were utilized for Cat ↔ Human face translations in conjunction with the CelebA-HQ dataset.

Evaluation Metrics. Following previous convention, two popular evaluation metrics, namely Fréchet Inception Distance (FID) [18] and Learned Perceptual Image Patch Similarity

TABLE II
QUANTITATIVE COMPARISON OF OUR VLC-UNIT WITH THE STATE-OF-THE-ART METHODS USING FID AND LPIPS.

Task	**Male ↔ Female**		**Dog ↔ Cat**		**Human Face ↔ Cat**		**Bird ↔ Dog**		**Bird ↔ Car**		**Average**	
Metric	**FID↓**	**LPIPS↑**	**FID↓**	**LPIPS↑**	**FID↓**	**LPIPS↑**	**FID↓**	**LPIPS↑**	**FID↓**	**LPIPS↑**	**FID↓**	**LPIPS↑**
TraVeLGAN [9]	66.60	–	58.91	–	85.28	–	169.98	–	164.28	–	109.01	–
U-GAT-IT [16]	29.47	–	38.31	–	110.57	–	178.23	–	194.05	–	110.12	–
MUNIT [2]	22.64	0.37	80.93	0.47	56.89	**0.53**	217.68	0.57	121.02	0.60	99.83	0.51
COCO-FUNIT [11]	39.19	0.35	97.08	0.08	236.90	0.33	30.27	0.51	207.92	0.12	122.27	0.28
StarGAN2 [12]	14.61	**0.45**	22.08	0.45	11.35	0.51	20.54	0.52	29.28	0.58	19.57	0.50
GP-UNIT [6]	14.63	0.37	15.29	0.51	13.04	0.49	11.29	0.60	13.93	0.61	13.64	**0.52**
PFC-UNIT [7]	11.55	0.35	9.30	0.52	10.21	0.50	9.21	0.60	11.26	0.62	10.30	0.52
Ours	**11.10**	0.36	**8.79**	**0.52**	**10.14**	0.51	**9.06**	**0.61**	**10.36**	**0.63**	**9.89**	**0.53**

(LPIPS) [19] were employed for quantitatively assessing the translation performance. The lower FID score indicates that the generated images are more realistic. The higher the LPIPS, the better the diversity of the generated images in the intra-domain.

Implementations. Our proposed VLC-UNIT was implemented using the PyTorch deep learning framework and the experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. All images are resized to 256 × 256 for both training and testing. For the text encoders used in VLC-UNIT, we adopted the pre-trained ViT-B/32 [20] from CLIP, with its weights fixed during the main training stage. Notably, the loss functions $\mathcal{L}_{adv}$, $\mathcal{L}_{sty}$, $\mathcal{L}_{con}$, and $\mathcal{L}_{msk}$ were used to train the generator over the initial 5000 iterations of the main training stage. Subsequently, the loss functions $\mathcal{L}_{gcla}$ and $\mathcal{L}_{rcla}$ were incorporated for further training. Typically, the discriminator's fourth layer is used to extract style features for $\mathcal{L}_{sty}$. When performing the task of translating human faces, the discriminator's fifth layer extracts style features for $\mathcal{L}_{sty}$. The weight parameters were set as follows: $\lambda_t = 50$, $\lambda_f = 0.25$, and $\lambda_q = 0.01$.

B. Comparison

This section presents both a quantitative and qualitative comparison of our proposed VLC-UNIT with state-of-the-art methods. Table. II provides a quantitative comparison across five tasks: Male ↔ Female, Dog ↔ Cat, Human Face ↔ Cat, Bird ↔ Dog, and Bird ↔ Car. These methods have also been reported in GP-UNIT [6]. Fig. 5 presents a visual analysis comparing our method with GP-UNIT, focusing on the consistency of style across fine-grained categories in the translation results.

For the Male ↔ Female task, our method achieves the best FID score of 11.10, significantly outperforming other methods. The LPIPS score of our method is 0.36, competitive with top-performing methods, with StarGAN2 achieving a slightly better score of 0.45. Visually, our proposed method excels in preserving the identity and distinct facial features of the original content while seamlessly integrating style attributes, resulting in realistic and well-defined images. In contrast, GP-UNIT outputs often exhibit a loss of identity features, leading to less coherent and realistic images.

For the Dog ↔ Cat task, our method also shows superior performance with the lowest FID score of 8.79 and a high LPIPS score of 0.52. This indicates that our approach generates more realistic images while maintaining high perceptual similarity. Visually, our method demonstrates superior capability in preserving the structural integrity and recognizable characteristics of the animals while effectively blending style elements. The generated images are visually coherent and lifelike. In contrast, GP-UNIT results are inconsistent in preserving content structure, often leading to distorted and less realistic images.

For the Human Face ↔ Cat task, our approach achieves an FID score of 10.14, close to the best performance by StarGAN2 (11.35). Our LPIPS score of 0.51 is also among the top results, demonstrating the robustness of our method in this challenging task.

For the Bird ↔ Dog task, our method achieves the lowest FID score of 9.06 and the highest LPIPS score of 0.61, indicating excellence in generating high-quality and perceptually similar images. Our approach maintains the core features of birds while adopting style attributes in a visually pleasing and coherent manner. The distinct bird features are well-preserved in the output images. Conversely, GP-UNIT often produces images with less detail, occasionally failing to retain essential features.

For the Bird ↔ Car task, our method performs exceptionally well with an FID score of 10.36 and an LPIPS score of 0.63, both the best among all compared methods. Visually, our method successfully transfers style attributes while preserving the structural integrity of cars, resulting in detailed and realistic outputs. In contrast, GP-UNIT outputs are less precise, with noticeable structural distortions that reduce realism.

On average, our method achieves the lowest FID score of 9.89 and the highest LPIPS score of 0.53 across all tasks. This demonstrates the superior performance and generalization ability of our method in generating high-quality and perceptually realistic images. Our proposed method consistently outperforms GP-UNIT in preserving content features while

TABLE III
ABLATION STUDY RESULTS FOR CAT → DOG TASK

Metric	w/o $\mathcal{L}_{sgcla}$	w/o $\mathcal{L}_{srcla}$	full model
FID	8.32	6.13	5.80
LPIPS	0.49	0.50	0.50

TABLE IV
ABLATION STUDY RESULTS FOR MALE → FEMALE TASK

Metric	w/o $\mathcal{L}_{crcla}$	w/o $\mathcal{L}_{cgcla}$	full model
FID	10.35	8.97	8.61
LPIPS	0.35	0.35	0.35

effectively integrating style attributes across diverse categories. The outputs from our method are more realistic, coherent, and detailed, highlighting its superior capability in handling various content and style inputs.

We conduct the ablation study to demonstrate the contribution of each component to the overall framework. Specifically, we evaluate the impact of the $\mathcal{L}_{sgcla}$ and $\mathcal{L}_{srcla}$ losses on the Cat → Dog task, as shown in Table III. Additionally, we assess the effects of the $\mathcal{L}_{cgcla}$ and $\mathcal{L}_{crcla}$ losses on the Male → Female task, as shown in Table IV. As shown in Table III, excluding $\mathcal{L}_{sgcla}$ results in a significantly higher FID score of 8.32, indicating less realistic images, while the LPIPS score slightly decreases to 0.49. Including $\mathcal{L}_{sgcla}$ but omitting $\mathcal{L}_{srcla}$ improves the FID score to 6.13 but remains inferior to the full model, which achieves a superior FID of 5.80 and consistent LPIPS of 0.50. Table IV shows that excluding $\mathcal{L}_{cgcla}$ results in a notably higher FID of 10.35, whereas retaining $\mathcal{L}_{cgcla}$ and removing $\mathcal{L}_{crcla}$ reduces the FID to 8.97, both falling short of the full model's 8.61. The LPIPS score remains constant at 0.35 across all configurations. These results underscore the critical role each proposed loss function plays in enhancing the realism and perceptual quality of the generated images, with the full model consistently delivering the best performance.

V. CONCLUSION

We found that current UNIT methods are significantly impacted by dataset distributions, leading to diminished performance on data-sparse instances and inconsistencies in fine-grained categories and target styles, thereby generating pseudo categories. In response to these challenges, we propose an innovative unsupervised image-to-image translation method called VLC-UNIT. By leveraging the vision-language classification capabilities of the CLIP model, VLC-UNIT uses text embeddings derived from multiple fine-grained detailed prompts to guide the image translation process. This approach ensures precise alignment of the translated outputs with the stylistic characteristics of the target style, enhancing both realism and stylistic diversity. Our method demonstrates superior performance over state-of-the-art techniques by effectively managing new categories without relying on extensive pre-training datasets, thus broadening the applicability of image translation technologies across various domains. Experimental results validate that VLC-UNIT significantly improves the quality and diversity of translated images, offering a robust solution for artistic output and data augmentation.

REFERENCES

[1] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022.

[2] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

[3] K. Baek, Y. Choi, Y. Uh, J. Yoo, and H. Shim, "Rethinking the truly unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14154–14163, 2021.

[4] J. Park, S. Kim, S. Kim, S. Cho, J. Yoo, Y. Uh, and S. Kim, "Lanit: Language-driven image-to-image translation for unlabeled data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23401–23411, June 2023.

[5] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–35, 2019.

[6] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Unsupervised image-to-image translation with generative prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18332–18341, 2022.

[7] Y.-Y. Liang and Y.-G. Wang, "Pfc-unit: Unsupervised image-to-image translation with pre-trained fine-grained classification," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1175–1179, 2023.

[8] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[9] M. Amodio and S. Krishnaswamy, "Travelgan: Image-to-image translation by transformation vector learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8983–8992, 2019.

[10] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–19, 2019.

[11] K. Saito, K. Saenko, and M.-Y. Liu, "Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 382–398, 2020.

[12] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197, 2020.

[13] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 11162–11173, 2021.

[14] M. B. Sariyildiz, J. Perez, and D. Larlus, "Learning visual representations with caption annotations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 153–170, 2020.

[15] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proceedings of the 7th Machine Learning for Healthcare Conference (MLHC)*, pp. 2–25, 2022.

[16] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–11, 2020.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–26, 2018.

[18] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.