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

In this document, we describe network architecture, details about domain descriptions, additional ablation study, additional experimental results, and user study for“LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data”.

Table of Contents

A. More Implementation Details

A.1. Network architecture of LANIT.

We summarize the detailed network architecture of our LANIT in Table 1. We basically follow the content encoder, style encoder, mapping network, and generator architecture from StarGAN2 [11].

Content Encoder

Layer	Resample	Norm	Output shape $(C \times H \times W)$
Conv1×1	-	-	$(64, 256, 256)$
Resblock	AvgPool	InstanceNorm	$(128, 128, 128)$
Resblock	AvgPool	InstanceNorm	$(256, 64, 64)$
Resblock	AvgPool	InstanceNorm	$(512, 32, 32)$
Resblock	AvgPool	InstanceNorm	$(512, 16, 16)$

Mapping Network

Layer	Type	Activation	Output shape (C)
Latent	Shared	-	16
Linear	Shared	ReLU	512
Linear	Shared	ReLU	512
Linear	Shared	ReLU	512
Linear	Shared	ReLU	512
Linear	Unshared	ReLU	512
Linear	Unshared	ReLU	512
Linear	Unshared	ReLU	512
Linear	Unshared	-	64

Generator

Layer	Resample	Norm	Output shape $(C \times H \times W)$
Resblock	-	InstanceNorm	$(512, 16, 16)$
Resblock	-	InstanceNorm	$(512, 16, 16)$
Resblock	-	AdaptiveInstanceNorm	$(512, 16, 16)$
Resblock	-	AdaptiveInstanceNorm	$(512, 16, 16)$
Resblock	Upsample	AdaptiveInstanceNorm	$(512, 32, 32)$
Resblock	Upsample	AdaptiveInstanceNorm	$(256, 64, 64)$
Resblock	Upsample	AdaptiveInstanceNorm	$(128, 128, 128)$
Resblock	Upsample	AdaptiveInstanceNorm	$(64, 256, 256)$
Conv1×1	-	-	$(3, 256, 256)$

Style Encoder and	**Discriminator**		
Layer	Type	Activation	Output shape $(C \times H \times W)$
Conv1×1	-	-	$(64, 256, 256)$
Resblock	AvgPool	InstanceNorm	$(128, 128, 128)$
Resblock	AvgPool	InstanceNorm	$(256, 64, 64)$
Resblock	AvgPool	InstanceNorm	$(512, 32, 32)$
Resblock	AvgPool	InstanceNorm	$(512, 16, 16)$
Resblock	AvgPool	-	$(512, 8, 8)$
Resblock	AvgPool	-	$(512, 4, 4)$
LReLU	-	-	$(512, 4, 4)$
Conv4×4	-	-	$(512, 1, 1)$
LReLU	-	-	$(512, 1, 1)$
Linear(Unshared)	-	-	$(64, 1, 1)$

Table 1. **Network architecture of our LANIT.**

A.2. Additional experimental setup.

We employ an Adam optimizer, where $\beta_1 = 0.0$ and $\beta_2 = 0.99$, for 100,000 iterations using a step decay learning rate scheduler. We also set a batch size of 8, an initial learning rate of 1e-4 for the encoder, generator, and discriminator, and 1e-6 for the prompt. All coefficients for the losses are set to 1. The training images are resized to 256×256. We conduct experiments using a single 24GB RTX 3090 GPU.

A.3. Template augmentation.

In addition, we utilize text-augmentation to boost the accuracy of pseudo labels. which are “a [domain] photo with [candidate domain].”, “a [domain] photo of the [candidate domain].”, “the [domain] photo of the [candidate domain].”, “a good [domain] photo of the [candidate domain].”, “high quality [domain] photo of [candidate domain].”, “a [domain] image of [candidate domain].”, “the [domain] image of [candidate domain].”, “high quality [domain] image of [candidate domain].”, “a high quality [domain] image of [candidate domain].” . The [domain] means the kind of text that represents the specific dataset such as “face”, “animal”, “food” etc. The [candidate domain] indicates the texts that can describe the specific domain as shown in Tab. 2

As shown in the Tab. 4, we can show that the template augmentation technique can boost the performance of pseudo labels.

A.4. Details of the latent-guided image-to-image translation.

As shown in Fig. 1, instead of utilizing reference images and a style encoder, we generated style vectors by inputting latent vectors sampled from a Gaussian distribution into the mapping network. Then style vectors were aggregated with pseudo labels provided by users.

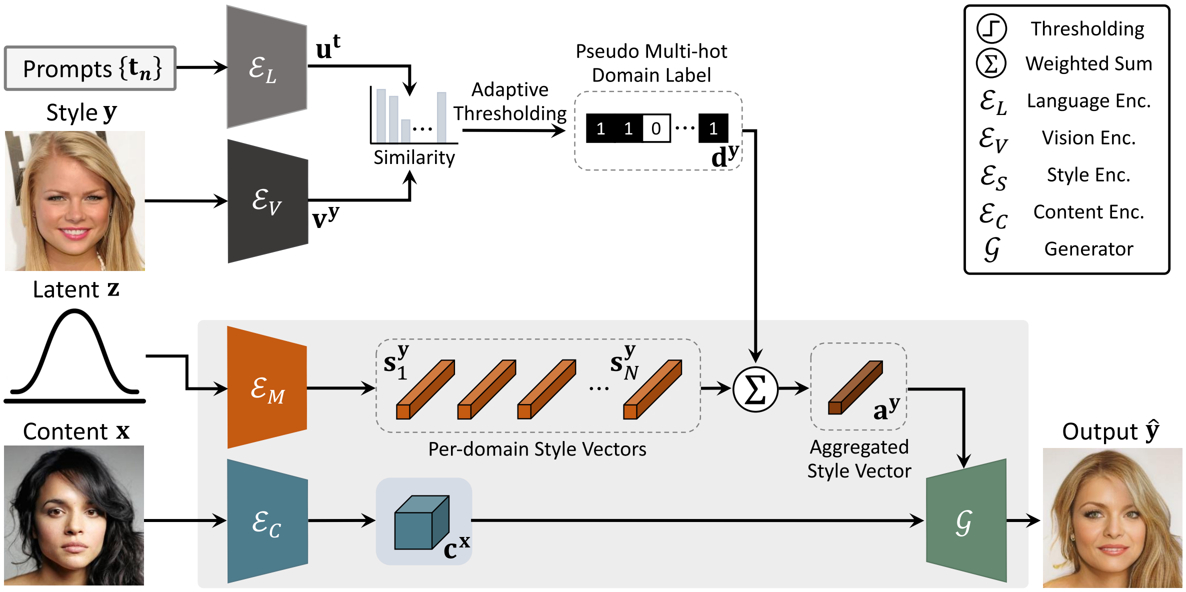


Figure 1. **Latent-guided image translation network.**

A.5. Details of templates and candidate domains.

Tab. 2 describes additional details of the candidate domains. We follow 10 pre-defined domains from TUNIT [3] for Animal Faces-10 and Food-10. Likewise, for CelebA-HQ [42], we obtain 40 pre-defined textual attributes from CelebA-HQ [42], and we have mainly shown the results using 10 domain descriptions, which are randomly selected for 3 times, and then report the average results. Please note that *we do not use per-sample domain labels* in all cases and the domain labels work as the dataset-level candidates.

Datasets	Template(default)	*N*	Candidate Domains
CelebA-HQ [42]	A face of	4	‘blond hair’, ‘bang’, ‘smiling’, ‘wearing lipstick’
		7	‘blond hair’, ‘black hair’ , ‘smiling’, ‘wearing lipstick’, ‘arched eyebrows’, ‘bangs’,‘mustache’
		10	‘bangs’, ‘blond hair’, ‘black hair’ ,‘smiling’, ‘arched eyebrows’, ‘heavy makeup’,‘mustache’, ‘straight hair’, ‘wearing lipstick’, ‘male’
Animal Faces-10 [36]	A photo of animal face with	4	‘beagle’, ‘golden retriever’, ‘tabby cat’, ‘tiger’
		7	‘beagle’, ‘dandie dinmont terrier’, ‘golden retriever’, ‘white fox’, ‘tabby cat’, ‘snow leopard’, ‘tiger’
		10	‘appenzeller sennenhund’, ‘beagle’, ‘dandie dinmont terrier’, ‘golden retriever’, ‘malinois’, ‘white fox’, ‘tabby cat’, ‘snow leopard’, ‘lion’, ‘tiger’
Food-10 [7]	A photo of food with	4	‘baby back ribs’, ‘beignets’, ‘dumplings’, ‘edamame’
		7	‘baby back ribs’, ‘beef carpaccio’, ‘beignets’, ‘clam chowder’, ‘dumplings’, ‘edamame’, ‘strawberry shortcake’
		10	‘baby back ribs’, ‘beef carpaccio’, ‘beignets’, ‘bibimbap’, ‘caesar salad’, ‘clam chowder’, ‘dumplings’, ‘edamame’, ‘spaghetti bolognese’, ‘strawberry shortcake’
LHQ [55]	A photo of scene	10	‘with mountain’, ‘with field’, ‘with lake’ ,‘with ocean’, ‘with waterfall’, ‘in summer’, ‘in winter’, ‘on sunny day’, ‘on cloudy day’, ‘at sunset’
MetFace	A portrait with	10	‘oil painting’, ‘grayscale’, ‘black hair’, ‘wavy hair’, ‘male’, ‘mustache’, ‘smiling’, ‘gray hair’, ‘blonde hair’, ‘sculpture’
Anime [9]	A photo of anime with	10	‘brown hair’, ‘red hair’, ‘black hair’, ‘purple hair’, ‘blond hair’, ‘blue hair’, ‘pink hair’, ‘silver hair’, ‘green hair’, ‘white hair’
LSUN-Car [62]	A car painted with	10	‘red color’, ‘orange color’, ‘gray color’, ‘blue color’, ‘yellow color’, ‘white color’, ‘black color’, ‘silver color’, ‘green color’, ‘pink color’
LSUN-Church [62]	A church	7	‘at night’, ‘with sunset’, ‘in winter’, ‘on cloudy day’, ‘on sunny day’, ‘with trees’, ‘with a river’

Table 2. **Examples of Templates and Candidate Domains for Each Dataset.**

A.6. Details of the predefined dictionary.

Tab. 3 describes additional details on candidate domains on a predefined dictionary. We used a pre-defined dictionary containing various domain descriptions for each dataset: CelebA-HQ, AnimalFaces-149, and Food-101 labels. As mentioned in the main paper, we selected 10 domains that have the highest similarity values within the whole image and target domains. We utilize the images of 10 classes used in the main paper and define the dictionary as the bundle of class names in the whole dataset. As can be seen in the table below, the selected candidate domains are mostly confident that have higher similarity values than other target domains.

Datasets	Selected Candidate Domains
CelebA-HQ [42]	'blond hair', 'bald', 'wavy hair', 'black hair', 'smiling', 'straight hair', 'eyeglasses', 'goatee', 'bangs', 'arched eyebrows'
Animal Faces-149 [36]	'dandie dinmont terrier','malinois','appenzeller sennenhund', 'white fox', 'tabby cat', 'snow leopard', 'lion', 'bengal tiger', 'grey fox', 'german shepherd dog'
Food-101 [7]	'french toast', 'beef carpaccio', 'beignets', 'seaweed salad', 'caesar salad', 'clam chowder', 'dumplings', 'edamame', 'spaghetti bolognese', 'strawberry shortcake'

Table 3. **Selected Candidate Domains in the Predefined Dictionary.**

B. Additional Ablation Study

B.1. Number of domain descriptions N.

In the main paper, we have examined the impacts of the number of candidate domains, base prompt thresholding, and prompt learning. In this section, we additionally validate the effects of the different number of domain descriptions, shown in Fig. 2 on CelebA-HQ. While our model consistently generates impressive outputs, results with a larger number of domain descriptions tend to faithfully represent diverse attributes. In addition, we evaluate the F1 score by varying the number of domains in Tab. 4. Our model shows the highest F1 score when N is 10 in both datasets.

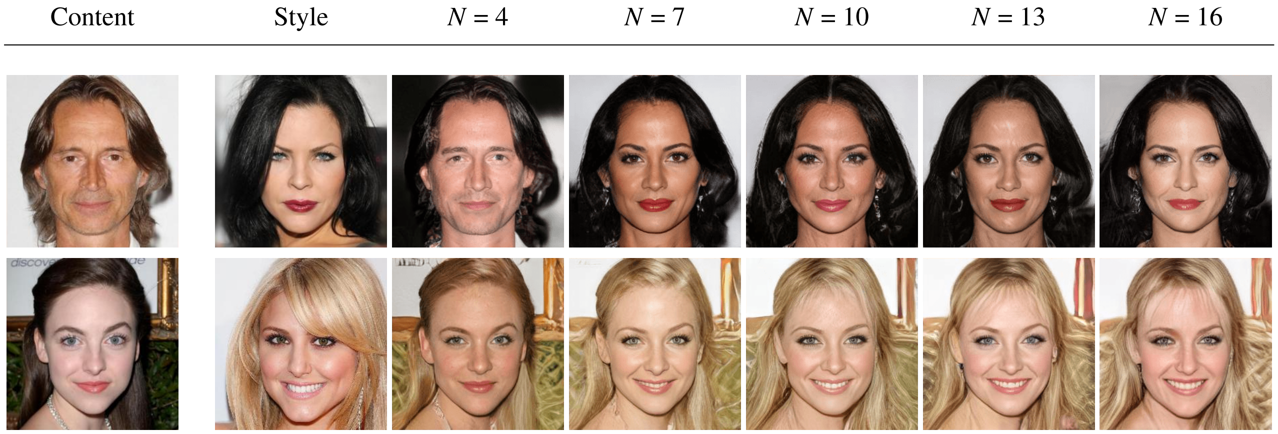


Figure 2. **Qualitative results by varying the number of domain description N.**

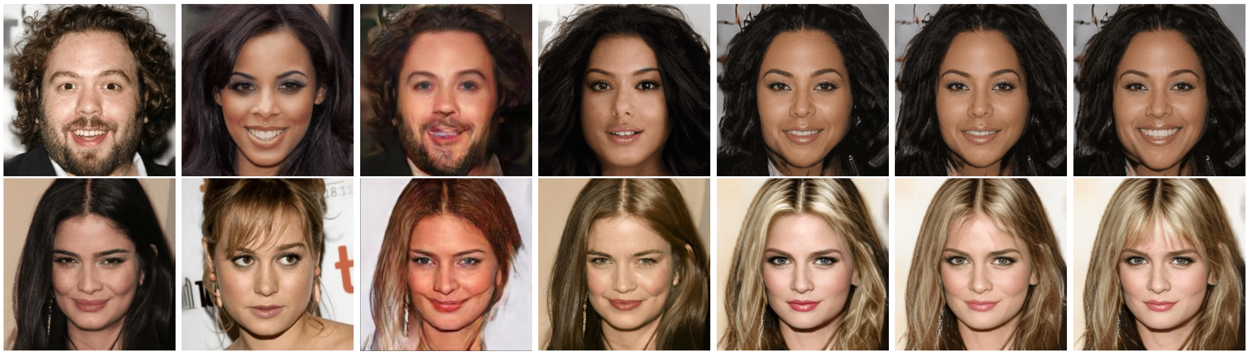
	AnimalFaces-10 [36]					CelebA-HQ [42]				
N	Top-1	Top-3	Baseline	TextAug	Prompt learning	Top-1	Top-3	Baseline	TextAug	Prompt learning
4	0.762	0.672	0.678	0.796	**0.832**	0.372	0.421	0.423	0.435	**0.481**
7	0.903	0.701	0.688	0.862	**0.893**	0.423	0.613	0.610	0.631	**0.652**
10	0.956	0.723	0.693	0.835	**0.880**	0.355	0.562	0.610	0.638	**0.670**
13	0.826	0.654	0.606	0.785	**0.801**	0.293	0.533	0.591	0.612	**0.639**
16	0.753	0.630	0.601	0.753	**0.783**	0.300	0.522	0.562	0.613	**0.641**

Table 4. **F1 score by varying number of domains.**

C. Additional Comparisons

C.1. Additional comparisons to other truly-unsupervised methods on reference-guided translation.

We additionally provide reference-guided image translation results, compared to existing fully-unsupervised I2I methods in Fig. 3: TUNIT [3] and Kim *et al.* [30]. Specifically, we visualize our LANIT with a different number of K (K=1,2,3). The results with more K have better capability to represent more diverse attributes. Meanwhile, our method is able to generate robust results with high fidelity regardless of K. On contrary, TUNIT and Kim *et al.* are limited to faithfully represent the style attributes from the reference images.

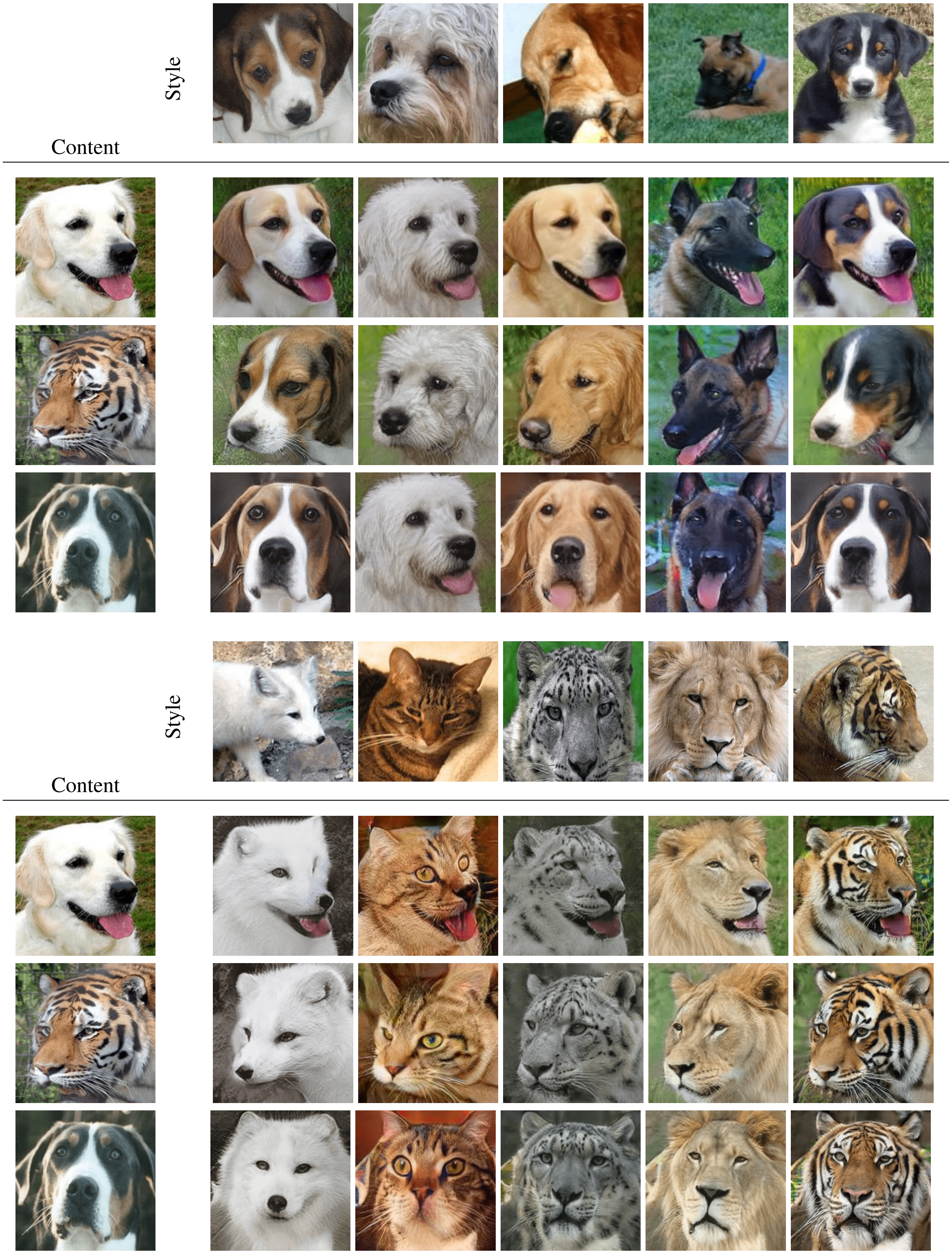


(a) Content (b) Style (c) TUNIT [3] (d) Kim *et al.* [30] (e) Ours(K=1) (f) Ours(K=2) (g) Ours(K=3)

Figure 3. **Reference-guided translation results.**

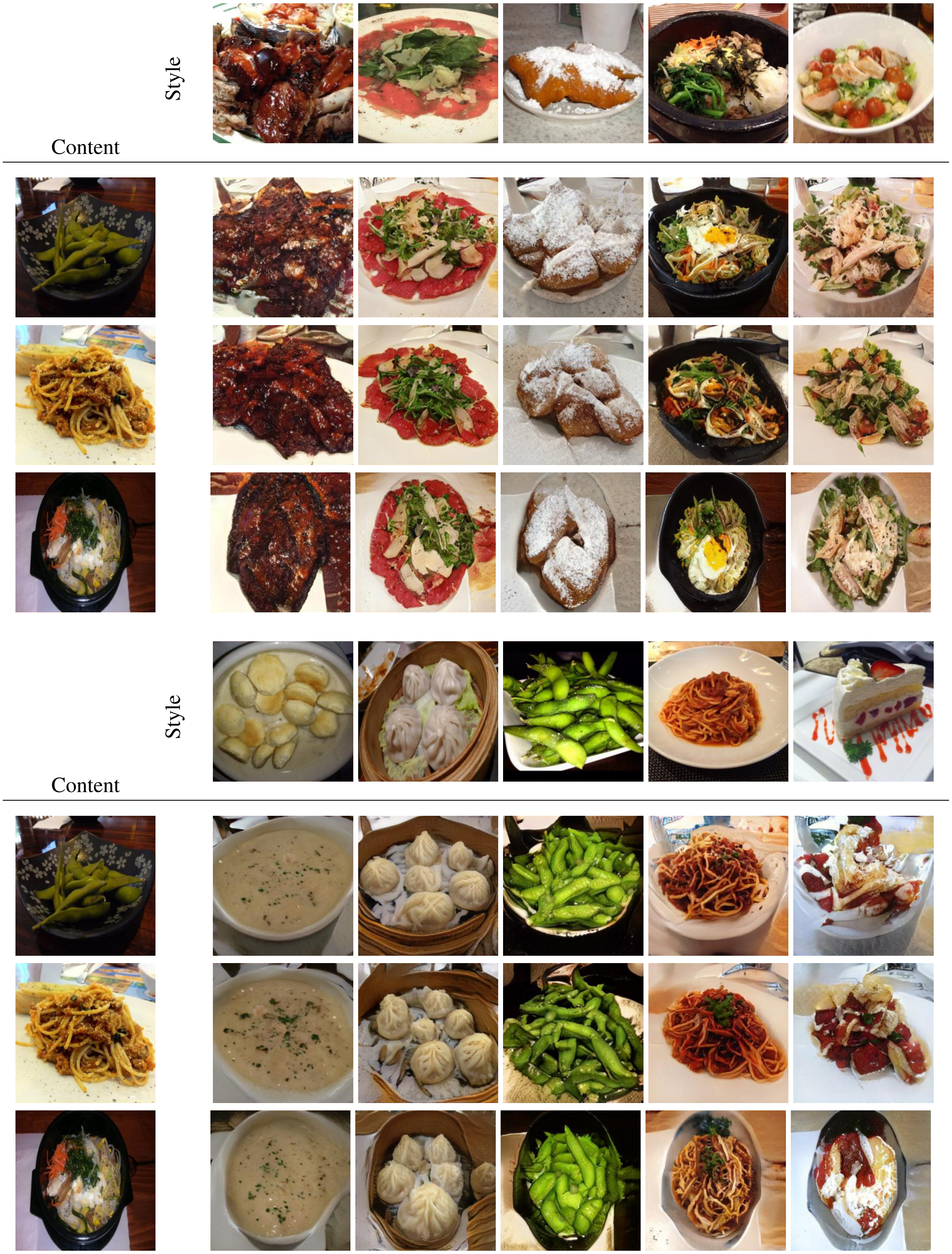
D. Additional Results of LANIT

In this section, we visualize additional visual results of our method in Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10 on Animal Faces-10, Food-10, CelebA-HQ, MetFace, Anime, LSUN-car and LSUN-church datasets [7,9,36,42,62], including latent-guided diverse image synthesis and reference-guided image translation results. Thanks to our multi-hot labeling setting with the proposed prompt learning technique and adaptive thresholding with the base prompt, our mapping network and style encoder can produce the style vectors faithfully representing target multiple domain styles.



Reference-guided image synthesis results

Figure 4. **Image translation results on Animal Faces-10.** Given domain descriptions are as follows: ‘appenzeller sennenhund’, ‘beagle’, ‘dandie dinmont terrier’, ‘golden retriever’, ‘malinois’, ‘white fox’, ‘tabby cat’, ‘snow leopard’, ‘lion’, ‘tiger’.



Reference-guided image synthesis results

Figure 5. **Image translation results on Food-10.** Given domain descriptions are as follows: ‘baby back ribs’, ‘beef carpaccio’, ‘beignets’, ‘bibimbap’, ‘caesar salad’, ‘clam chowder’, ‘dumplings’, ‘edamame’, ‘spaghetti bolognese’, ‘strawberry shortcake’.

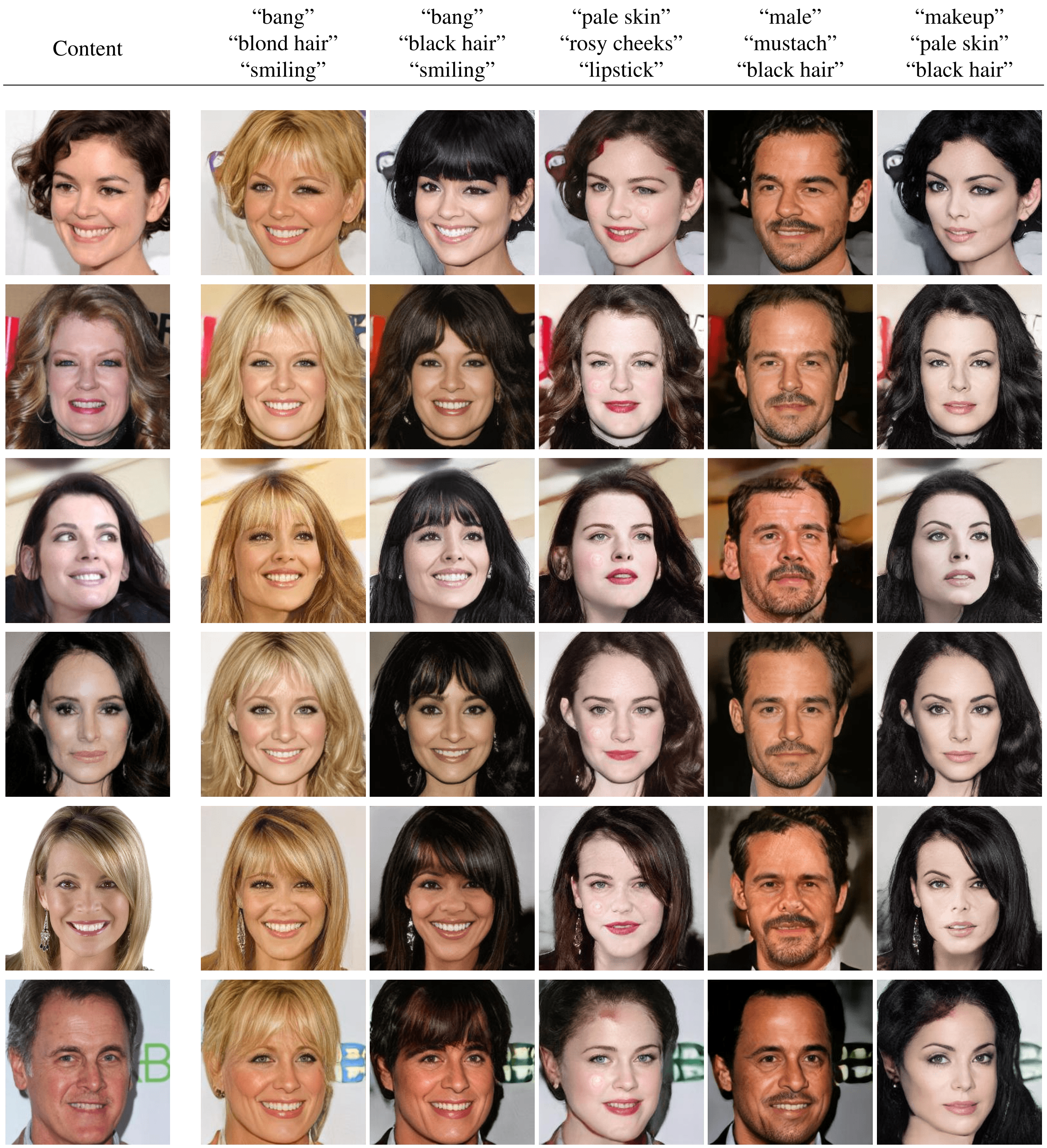


Figure 6. **Latent-guided diverse image synthesis results by our LANIT on CelebA-HQ.** Given domain descriptions are as follows: ‘bangs’, ‘blond hair’, ‘black hair’,‘smiling’, ‘pale skin’, ‘heavy makeup’, ‘no beard’, ‘rosy cheeks’, ‘wearing lipstick’, ‘male’.

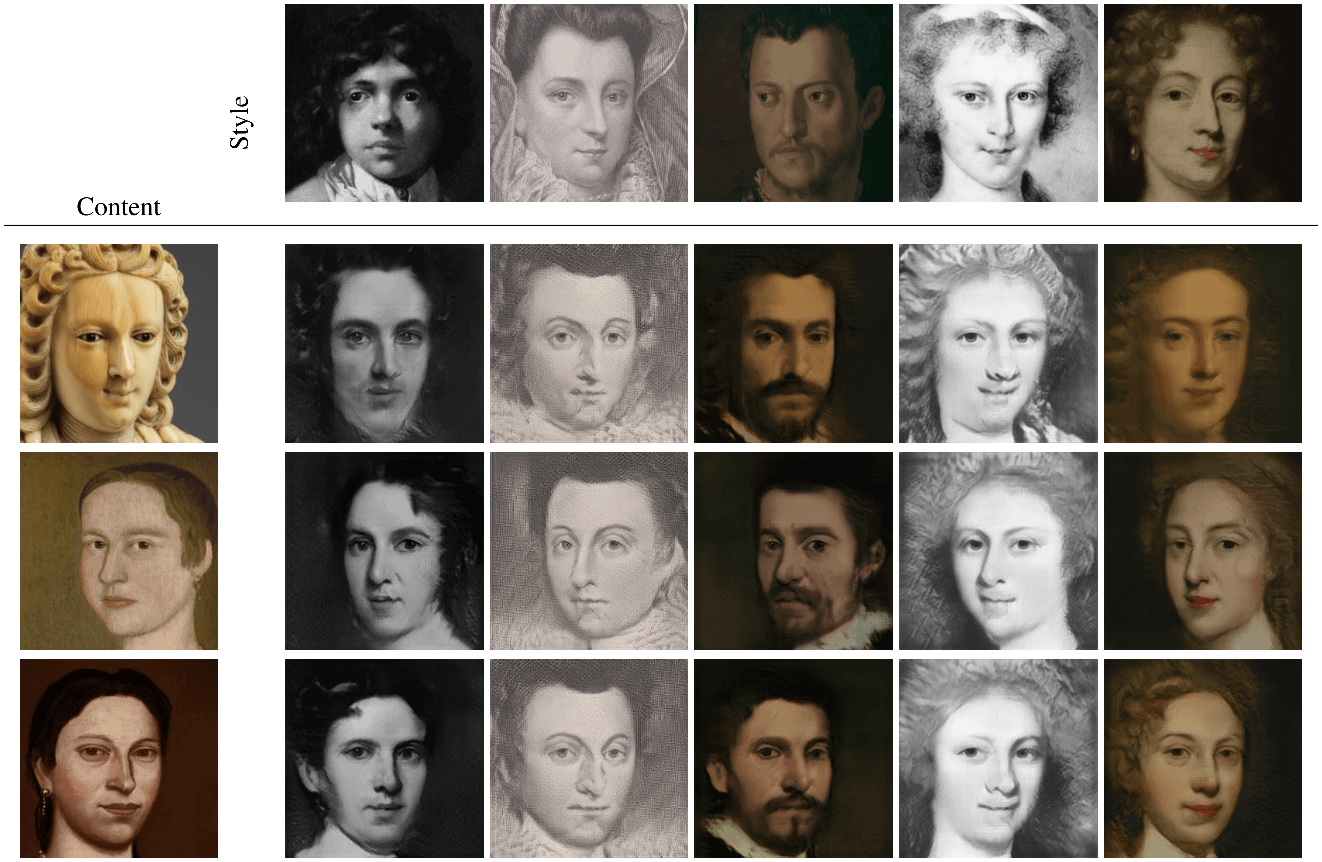


Figure 7. **Reference-guided image translation results by our LANIT on MetFace.** Given domain descriptions are as follows: “oil painting”, “grayscale”, “black hair”, “wavy hair”, “male”, “mustache”, “smiling”, “gray hair”, “blonde hair”, “sculpture”.

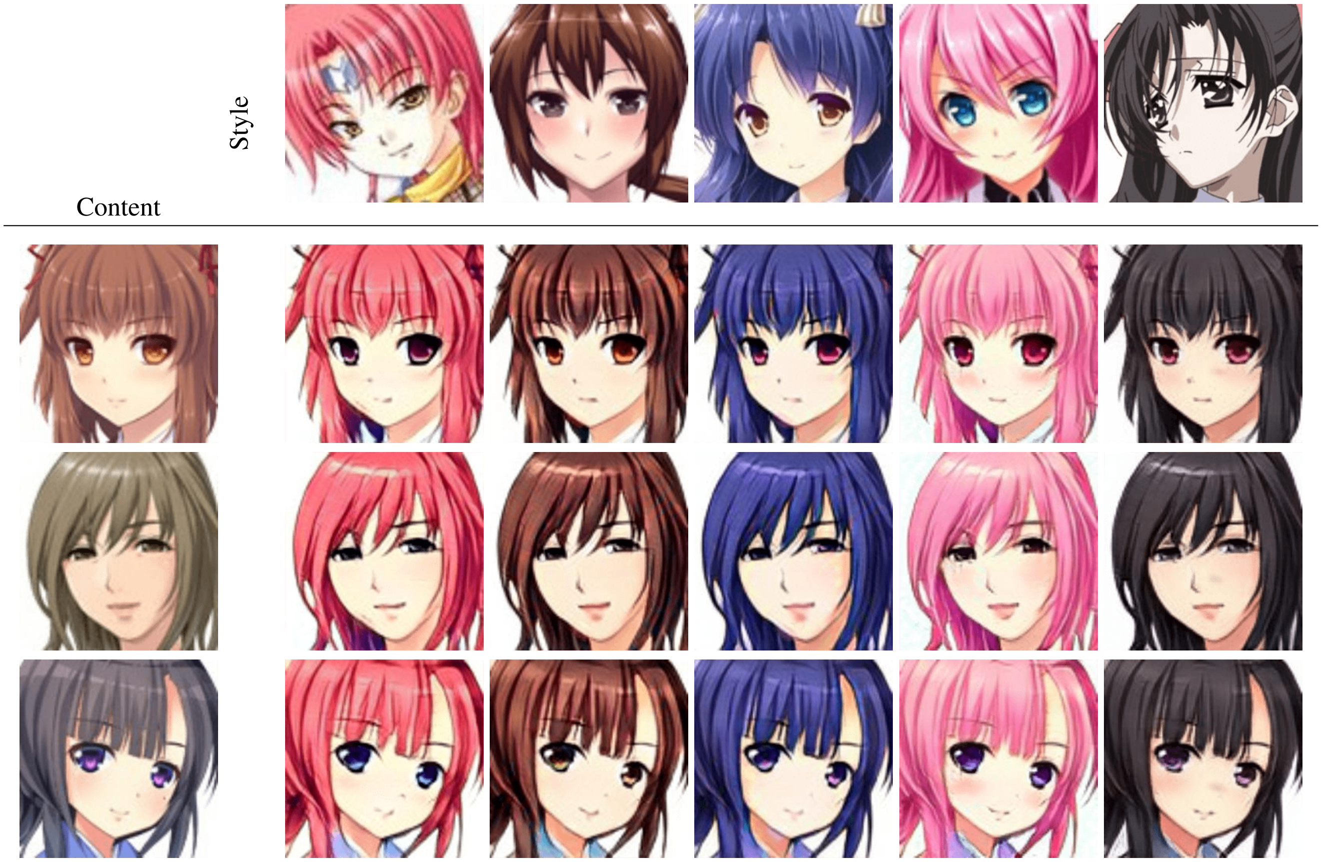


Figure 8. **Reference-guided image translation results by our LANIT on Anime [9].** Given domain descriptions are as follows: “brown hair”, “red hair”, “black hair”, “purple hair”, “blond hair”, “blue hair”, “pink hair”, “silver hair”, “green hair”, “white hair”.

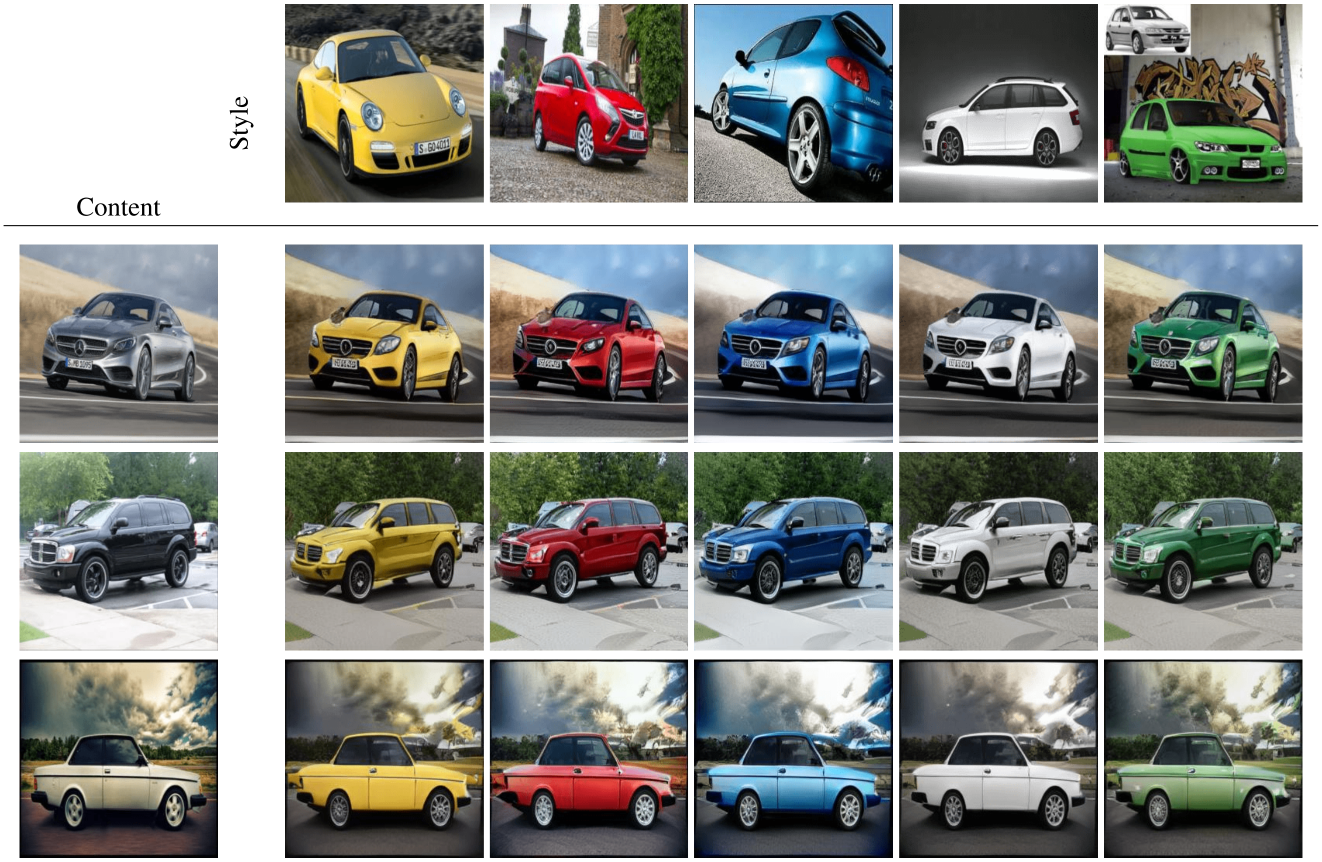


Figure 9. **Reference-guided image translation results by our LANIT on LSUN-car.** Given domain descriptions are as follows: “red color”, “orange color”, ‘gray color”, “blue color”, “yellow color”, “white color”, “black color”, “silver color”, “green color”, “pink color”.

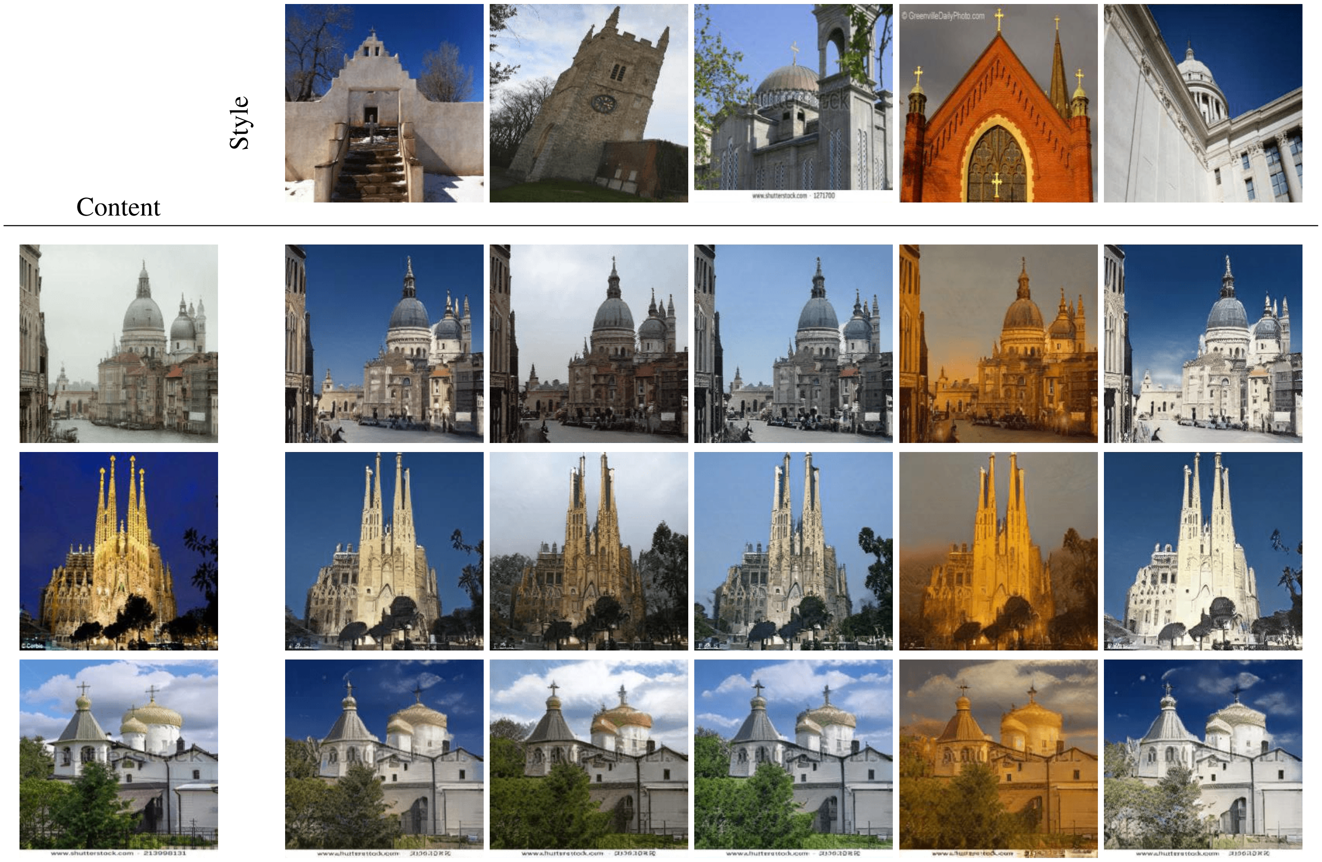


Figure 10. **Reference-guided image translation results by our LANIT on LSUN-church.** Given domain descriptions are as follows: “at night”, “at sunset”, “in winter”, “on cloudy day”, “on sunny day”, “with trees”, “with a river”.

E. User Study.

Finally, we conducted a user study to evaluate the image quality, content preservation, and style consistency of LANIT compared to StarGAN2, TUNIT, and Kim *et al*. 204 users were involved in this study, asked to answer 60 questions, each of which is generated from randomly sampled 20 images from CelebA-HQ. The examples of questions are as follows: "Which results do you think have the highest quality/preserve content information such as pose/have similar style of the reference image?" Fig. 11 shows our LANIT achieves the top rank in all tasks.

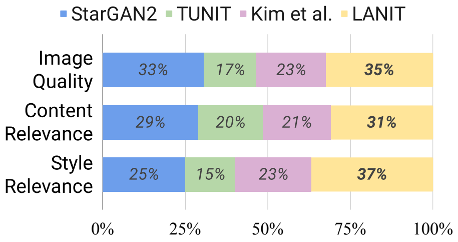


Figure 11. **User study results.**

F. Limitations

Although our LANIT shows outstanding performance on various benchmarks, LANIT inherits some problems from pre-trained vision-language models. In specific, to overcome this, we suggest adaptive thresholding and prompt learning techniques and these techniques effectively boost the confidence of pseudo labels. However, these techniques still have limitations for getting accurate pseudo labels, which degrades the performance of the image-to-image translation framework.