Supplementary Material: Smoothing the Disentangled Latent Style Space for Unsupervised Image-to-Image Translation

Yahui Liu[1,3], Enver Sangineto[1], Yajing Chen[2], Linchao Bao[2], Haoxian Zhang[2], Nicu Sebe[1], Bruno Lepri[3], Wei Wang[1*], Marco De Nadai[3*]

[1]University of Trento, Italy [2]Tencent AI Lab, China [3]Fondazione Bruno Kessler, Italy

A. Model Architecture

Fig. 1 shows the framework of our proposed method for MMUIT tasks. The model is composed of an image generator G, a discriminator D, an encoder E and an MLP M. G generates a new image from a source image $\boldsymbol{x}_a$ and a style code $\boldsymbol{s}$, which can either be extracted from a reference image (i.e. $\boldsymbol{s}_p = E(\boldsymbol{x}_p)$, or from a randomly sampled vector $\boldsymbol{z} \sim N(0,1)$ through $\boldsymbol{s}_p = M(\boldsymbol{z})$. The discriminator D learns to classify an image as either a real image in its associated domain, or a fake image.

As explained in the main paper, we use $\mathcal{L}_{tri}$, $\mathcal{L}_{SR}$ and $\mathcal{L}_{cont}$ to compact and disentangle the style space and to help preserving the source content. In Fig. 1, $\boldsymbol{s}_n$ is a style code of a domain different from the domain shared by $\boldsymbol{s}_p$ and $\boldsymbol{s}_a$.

B. Analysing the Style-Space Compactness

B.1. Inter-domain Distance Distributions

In order to estimate the inter-domain distances and the degree of compactness of a high-dimensional semantic space, we compute the distribution of the distances $(d_s(\boldsymbol{s}_a, \boldsymbol{s}_n) - d_s(\boldsymbol{s}_a, \boldsymbol{s}_p))$. Specifically, we use the CelebA-HQ dataset [4] and we randomly sample 10,000 triplets $(\boldsymbol{s}_a, \boldsymbol{s}_p, \boldsymbol{s}_n)$ where $\boldsymbol{s}_a \sim \boldsymbol{\mathcal{S}}_i$, $\boldsymbol{s}_p \sim \boldsymbol{\mathcal{S}}_i$ and $\boldsymbol{s}_n \sim \boldsymbol{\mathcal{S}}_j$ with $i \neq j$. Fig. 2 shows the distribution of $(d_s(\boldsymbol{s}_a, \boldsymbol{s}_n) - d_s(\boldsymbol{s}_a, \boldsymbol{s}_p))$ under different experimental settings.

Fig. 2 (a) shows that the distance distribution of the baseline system (without using $\mathcal{L}_{tri}$ and $\mathcal{L}_{SR}$) is relatively wide and corresponds to the largest median. Our $\mathcal{L}_{tri}$ loss with a small margin can slightly reduce both the range between the lower quartile to upper quartile and the range between the minimum to the maximum score. Conversely, $\mathcal{L}_{SR}$ ($\lambda_{SR} = 1.0$) compacts the space significantly. Jointly using $\mathcal{L}_{SR}$ and $\mathcal{L}_{tri}$ ($\alpha = 0.1$), the $\mathcal{L}_{SR}$-only distribution is slightly shifted up. Fig. 2 (b) shows the impact of λ_{SR} when we use $\mathcal{L}_{SR}$ without $\mathcal{L}_{tri}$. Conversely, Fig. 2 (c) analyses the case of jointly using $\mathcal{L}_{SR}$ (with $\lambda_{SR} = 1.0$) and $\mathcal{L}_{tri}$ while changing the margin α. The latter experiment shows that the Triplet Margin loss can adjust the distance between style clusters, since the ranges between the minimum and the maximum score are shifted when using a larger α.

The corresponding PS scores are presented in Fig. 3, which shows that increasing λ_{SR} helps smoothing the space, but when $\lambda_{SR} > 0.5$, only limited improvements are obtained (see Fig. 3 (a)).

As shown in the main paper, the Triplet loss significantly influences the image quality and smoothness of I2I translations. Interestingly, the margin α also plays an important role. Using a small positive margin (e.g., 0.1) is enough to keep the disentanglement and achieve the best PS score, as shown in Fig. 3 (b). Meanwhile, a large margin can push the style clusters far away from each other, which may be harmful for the smoothness degree of the space.

B.2. An Alternative Style Regularization

A possible alternative to the style-regularization loss ($\mathcal{L}_{SR}$), is based on the following formulation, whose goal is to compact the style codes close to the surface of the zero-centered, n-dimensional unit sphere:

$$\mathcal{L}_{sph} = \mathbb{E}_{\boldsymbol{s} \sim \boldsymbol{\mathcal{S}}} \left[| \|\boldsymbol{s}\|_2 - 1 | \right] \tag{1}$$

where $\| \cdot \|_2$ is the L_2 norm. Note that, since the volume of the whole n-sphere is larger than the volume of its surface, $\mathcal{L}_{sph}$ leads to a much more compact space compared to $\mathcal{L}_{SR}$. Tab. 1 quantitatively compares $\mathcal{L}_{sph}$ with $\mathcal{L}_{SR}$ and shows that a very compact space ($\mathcal{L}_{sph}$) leads to a higher smoothness but with a low diversity. This finding is qualitatively confirmed in Fig. 4. This comparison indicates that there exists a trade-off between the smoothness of the space and the diversity of generated images.

B.3. A Space Visualization Experiment

We perform an additional experiment on the MNIST dataset [7] to interpret the results of our model and directly visualize the distributions of style codes. In this experiment, we consider the categories of handwritten digits

*Both authors contributed equally to this work.

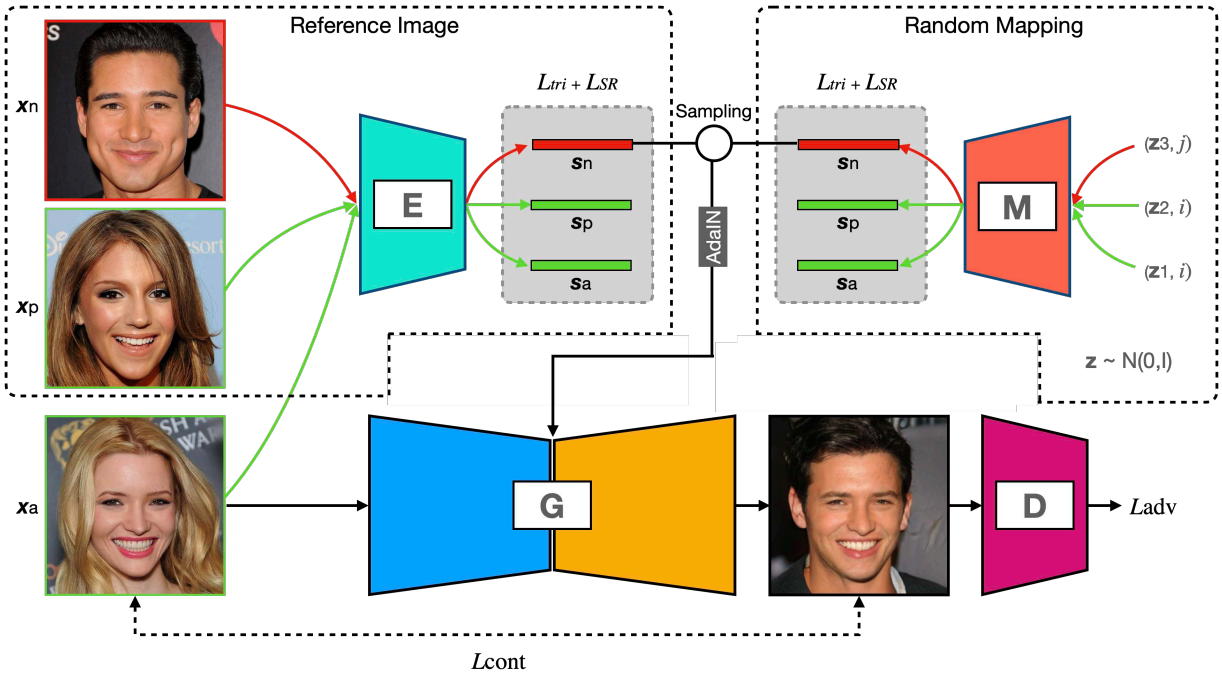
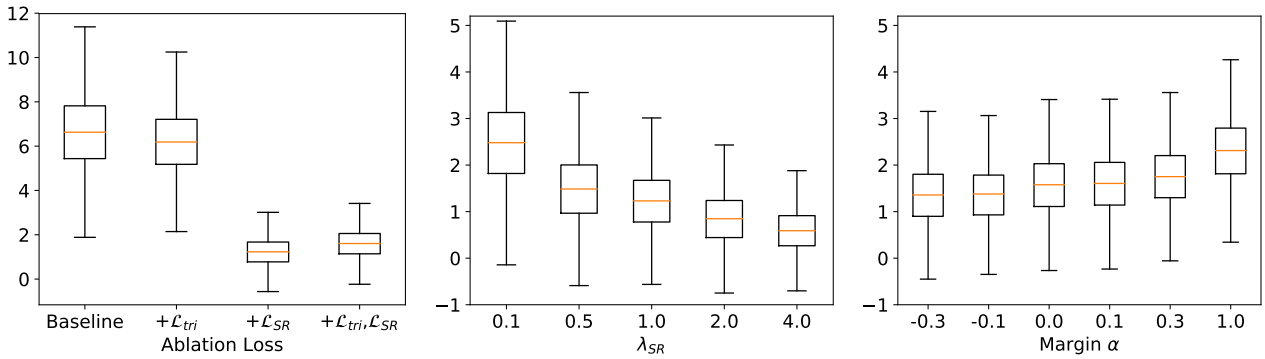


Figure 1: Our MMUIT generative framework and the style-code sampling strategies.



(a) (b) (c)

Figure 2: Distribution of $(d_s(\mathbf{s}_a, \mathbf{s}_n) - d_s(\mathbf{s}_a, \mathbf{s}_p))$ on different experimental settings on the CelebA-HQ dataset. (a) shows that $\mathcal{L}_{SR}$ helps to compact the style space, while $\mathcal{L}_{tri}$ can adjust the distance between the style clusters. (b) shows that the weight of the $\mathcal{L}_{SR}$ can control the compactness of the style space. (c) shows that increasing the margin α in $\mathcal{L}_{tri}$ has an effect on the distances between clusters.

Model	FID↓	LPIPS↑	PS↑	FRD↓
$\mathcal{L}_{SR}$	**23.37**	**.337**	.504	.837
$\mathcal{L}_{sph}$	23.66	.103	**.897**	**.808**

Table 1: A comparisons between $\mathcal{L}_{SR}$ and $\mathcal{L}_{sph}$ on a gender translation task using the CelebA-HQ dataset.

as "styles" and we set the dimension of style codes to 2, such that they can be easily plotted in a two-dimensional coordinate system without reducing the representation dimensionality with non-linear projections (e.g. t-SNE). As shown in Fig. 5 (a), the original style codes without using our proposed losses, is scattered in a non-compact space, where there are many "training gaps". Once we increase the weight of λ_{SR}, the style codes are pushed in a more compact space. However, the clusters (i.e., the domains) are highly entangled, as shown in Fig. 5 (b). Conversely, the triplet loss alleviates this issue by separating the compacted clusters, as shown in Fig. 5 (c).

Moreover, we select two clusters with large "training gaps" (i.e., "2" (green color) and "7" (grey color)) in the original space Fig. 5 (a). Fig. 6 (a) shows an example of interpolation results between "2" and "7" with large "training gaps", showing, as expected, that the generated images

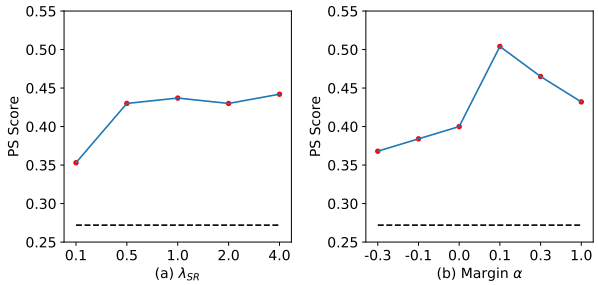


Figure 3: An ablation study on the influence of both (a) the SR loss weigh λ_{SR} and (b) the triplet loss margin α ($\lambda_{SR} = 1.0$) in the PS scores. The black dashed line refers to StarGAN v2 [3].

contain artifacts. Fig. 6 (b) refers to the same interpolation between "2" and "7" in the setting with $\lambda_{SR} = 1.0$. It seems that, due to the cluster overlapping, the interpolation traverses another cluster (i.e., "4") while moving from "2" to "7". Finally, the triplet loss is able to disentangle the compact space, as shown in Fig. 6 (c), where no "intruder" is generated when interpolating between the two domains.

C. PS Details

The proposed PS score requires a perceptual distance metric $\psi(\cdot,\cdot)$. We chose to use the LPIPS [11] distance, which was shown to well align to human judgements. However, although Zhang et al. [11] claim that LPIPS is a metric, its formulation is based on the squared Euclidean distance between deep learning features:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} w_l \|\boldsymbol{y}_1^l - \boldsymbol{y}_2^l\|_2^2 \qquad (2)$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are image patches, $\boldsymbol{y}^l$ is a feature extracted with a pretrained network $\mathcal{F}$ (e.g., AlexNet [6]) using its l-th layer, and the weights w_l are layer-specific weights trained to mimic the human perception. Thus, Eq.(2) does not obey to the Triangle Inequality, which is necessary for a distance to be a metric. To avoid this problem, we re-train the w_l weights using an Euclidean-distance formulation, which gives us a proper metric (called LPIPS* in the rest of this Supplementary Material):

$$d'(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} w_l \|\boldsymbol{y}_1^l - \boldsymbol{y}_2^l\|_2. \qquad (3)$$

Following the original paper [11], the network $\mathcal{F}$ used in our paper is an AlexNet [6] pre-trainted on ImageNet where a linear classifier (i.e., the the w_l weights) is trained to learn a human perception distance.

Comparison with other smoothness metrics. The smoothness of a latent style space can also be evaluated using LPIPS [11] and the PPL [5] scores. In ideally smooth interpolations, the perceptual distance (LPIPS) between two neighbouring interpolations should be as low as possible (i.e., high similarity). Similarly, the PPL should be as low as possible to indicate the smoothness of the space. Note that when the model exhibits a mode collapse problem, we can have PPL=0 (or LPIPS=0). Despite this, we compare the LPIPS, PPL and PS scores on an additional experiment, where we randomly use both intra and inter-domain interpolation lines. For each interpolation line we generate 20 images. Tab. 2 shows that: (1) the higher the PS score, usually the lower the LPIPS and the PPL score; (2) our PS metric based on LPIPS* is more consistent with the LPIPS and the PPL with respect to the smoothness degree. Moreover, our PS metric is more interpretable, as it ranges between 0 and 1, while the alternatives range in $[0, \infty]$.

Model	Percep. Distance	PS↑ Intra	PS↑ Inter	LPIPS↓ Intra	LPIPS↓ Inter	PPL↓ Intra	PPL↓ Inter
[3]	LPIPS	.877	.670	.005	.012	19.21	57.19
	LPIPS*	.545	.359	.061	.107		
Ours	LPIPS	.850	.840	.003	.006	9.84	22.78
	LPIPS*	.625	.485	.047	.071		

Table 2: Comparing different smoothness metrics. We use two different basic perceptual distances for all the metrics: the original LPIPS (Eq.(2)) and the revised LPIPS* (Eq.(3)). The LPIPS column refers to the diversity degree [11]. "Intra" and "Inter" refer to intra-domain and inter-domain interpolations, respectively.

Percepual Distance	Num. of Interpolation 10	20	50	100
PPL	120.63	457.53	2122.33	6369.93
LPIPS	0.133	0.106	0.066	0.042
LPIPS*	1.150	1.424	1.723	1.908

Table 3: The sum of the perceptual distances along the same interpolation lines averaged over all the generated images. This table shows the linearity of various perceptual distance metrics.

Number of Interpolations. We also compute the robustness of the different metrics on a high number of interpolations in Tab. 3, where we use the same start-end style codes for all the metrics. Tab. 3 shows that PPL is not a linear metric and it is sensitive to the interpolation step size (i.e., the smaller the interpolation step size, the larger the PPL score). Similarly, LPIPS is also not a linear metric and it tends to decrease when the number of interpolations increase. Conversely, the proposed PS score is consistent, it satisfies the triangle inequality and its behaviour is more linear.

Figure 4: Visual comparisons between (a) $\mathcal{L}_{SR}$ and (b) $\mathcal{L}_{sph}$.

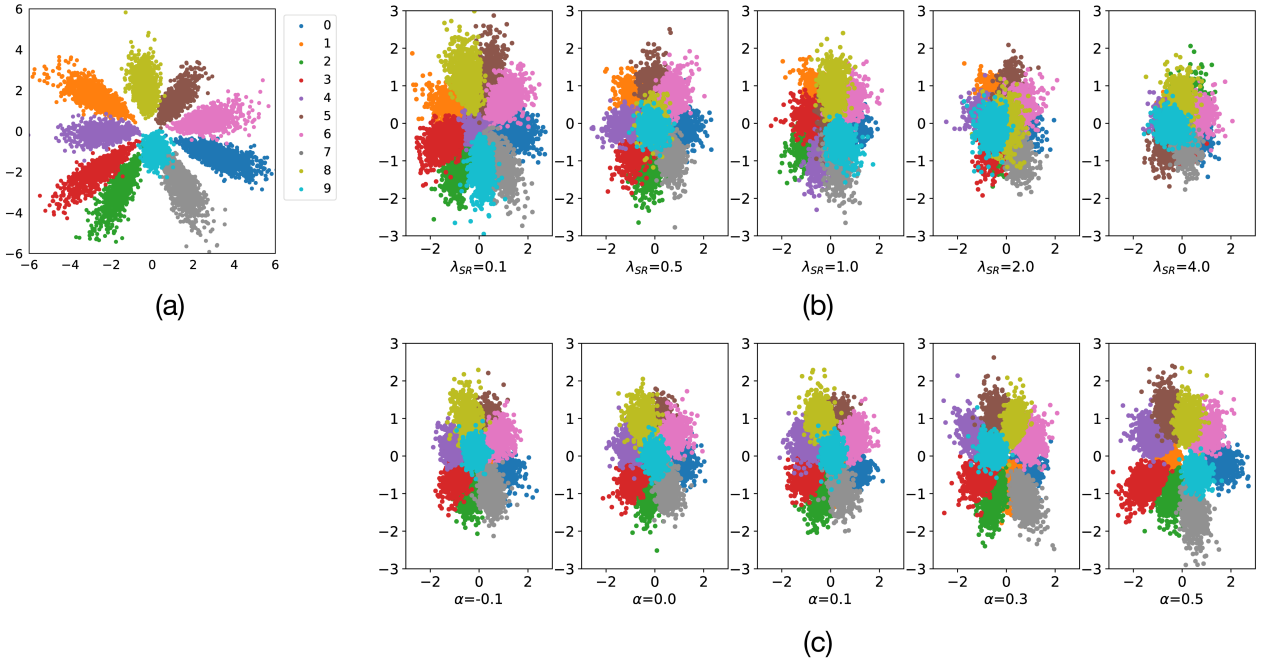


Figure 5: The distributions of style codes on a MNIST-based toy experiment. The original latent style space (a), using only $\mathcal{L}_{SR}$ with different loss weights λ_{SR} (b), and using $\mathcal{L}_{SR}$ ($\lambda_{SR} = 1.0$) and $\mathcal{L}_{tri}$ with different margin values α (c).

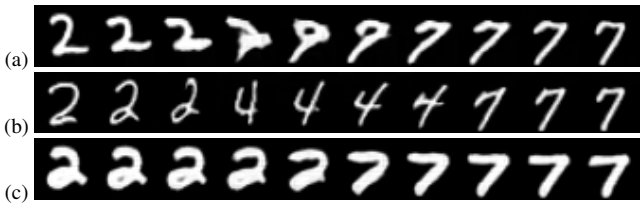


Figure 6: Interpolations results on MNIST between domain "2" and domain "7". (a) Original space, (b) Using only $\mathcal{L}_{SR}$ ($\lambda_{SR} = 1.0$). (c) Using $\mathcal{L}_{SR}$ ($\lambda_{SR} = 1.0$) and $\mathcal{L}_{tri}$ ($\alpha = 0.5$).

Model	Interpolation	PS↑		LPIPS↓		PPL↓	
		Intra	Inter	Intra	Inter	Intra	Inter
[3]	Lerp	.545	.359	.061	.107	19.21	57.19
Ours	Lerp	**.625**	**.485**	**.047**	**.071**	**9.84**	**22.78**
[3]	Slerp	.531	.336	.065	.120	19.69	64.81
Ours	Slerp	**.607**	**.404**	**.049**	**.083**	**10.53**	**26.17**

Table 4: Different interpolation strategies. Both StarGAN v2 [3] and the our method achieve a better performance with "lerp".

Interpolation Strategies. Finally, we test the robustness of the PS score with respect to two different interpolation strategies (i.e., *lerp* and *slerp* [5]). As shown in Tab. 4, both our methdod and StarGAN v2 [3] achieve a slightly better result when using the linear interpolation (*lerp*), which indicates the linearity of the style space.

D. Face Recognition Distance

Fig. 7 shows an example of face translation, which indicates the crucial issue of identity preservation. For example, an arbitrary female face can be realistic for a discriminator, but if the original-person identity is completely lost, this is not the desired output of a gender translation. Fig. 8 shows a comparisons based on a smile translations task on the CelebA-HQ dataset, which further shows the importance of the identity preservation. The StarGAN generated images frequently loose the identity of the source images, while ours do not. Moreover, we see that $\mathcal{L}_{cont}$ is very important both for the identity and the background preservation.

E. LPIPS for Diversity

The state of the art models are often evaluated through the LPIPS distance. Usually, for each input, different target styles are randomly sampled. Then, the LPIPS is computed on all the generated outputs to model the diversity (also called multi-modality) of the generated images. However, a high LPIPS distance is not always desirable. For example, a high LPIPS value can be produced also when:

- The generated images do not always look real (e.g. the images with artifacts shown in the first row of Fig. 9).
- The domain-independent part of the image is not preserved. For example, when the background appearance has drastically changed (e.g., Fig. 8 (a)) or when the person-identity is not preserved (e.g., Fig. 8 (a) and Fig. 8 (b)).

For these reasons, we believe that in an MMUIT task, LPIPS scores should be taken with a pinch of salt, especially when the model is not good enough to preserve the domain-independent part of the source image.

F. Additional Details

F.1. Datasets

Following StarGAN v2 [3], we use the CelebA-HQ [4] and the AFHQ [3] dataset. CelebA-HQ is a high-quality version of the CelebA [9] dataset, consisting of 30,000 images with a 1024×1024 resolution. We randomly select 2,000 images for testing and we use all the remaining images for training. Differently from StarGAN v2, we also test the smile and the age attributes. AFHQ consists of 15,000 high-quality images at 512×512 resolution. The dataset includes three domains (cat, dog, and wildlife), with 5,000 images each. We select 500 images as the test set for each domain and we use all the remaining images for training. AFHQ and CelebA-HQ are tested at a 256×256 resolution (note that we use a 128×128 resolution in the comparisons with TUNIT [1]). In this Supplementary Material we also used the low-resolution MNIST [8] dataset, which consists of 60,000 training samples and 10,000 testing samples with a 32×32 resolution.

F.2. Compared Methods

We use the official released codes for all the compared methods, including StarGAN v2 [3][1], HomoGAN [2][2], InterFaceGAN [10][3] and TUNIT [1][4]. In the main paper (Sec. 4.2) we show how our proposed losses are combined with (i.e., simply added to) the StarGAN v2 losses. Similarly, in case of TUNIT, we use all the original losses of [1] ($\mathcal{L}_{tunit}$) and we add $\mathcal{L}_{SR}$ and $\mathcal{L}_{tri}$ (without using our content loss), which leads to: $\mathcal{L}_{tunit} + \mathcal{L}_{SR} + \mathcal{L}_{tri}$.

InterFaceGAN [10] is not a I2I translation model, and there is no separation between the "content" and the "style" representations. Moreover, this method linearly interpolates codes on a StyleGAN [5] pre-trained semantic space. Thus, it is not easy to fairly compare MMUIT models with InterFaceGAN. In our paper, when we compare MMUIT models with InterFaceGAN, we start from a StyleGAN generated image $\boldsymbol{x}$ and we modify its semantics by generating two new images $\boldsymbol{x}' = G(\boldsymbol{z} + -3\boldsymbol{n}))$ and $\boldsymbol{x}'' = G(\boldsymbol{z} + 3\boldsymbol{n}))$, where $\boldsymbol{n}$ is the unit normal vector defining a domain-separation hyperplane (e.g. smile vs non-smile) learned by InterFaceGAN. In the semantic space of smile, $\boldsymbol{x}'$ is an image with *no smile*, while $\boldsymbol{x}''$ an image with *more smile*. These two randomly images are then used as the reference images for the encoders of each compared model (including ours) to generate the style codes. Note that, using StyleGAN based reference images, most likely favours InterFaceGAN with respect to all the other compared methods.

G. More Experiments

More visual comparisons with StarGAN v2 [3], HomoGAN [2], InterFaceGAN [10] and TUNIT [1] are shown in Fig. 9-10. Fig. 11-13 show more visual results of gender, smile and age translations on the CelebA-HQ dataset. Fig. 14-16 show more visual results of animal translations on the AFHQ dataset.

[1] https://github.com/clovaai/stargan-v2
[2] https://github.com/yingcong/HomoInterpGAN
[3] https://github.com/genforce/interfacegan
[4] https://github.com/clovaai/tunit

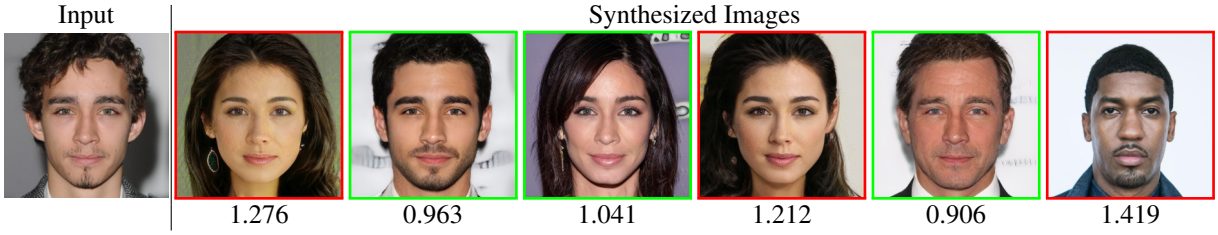


Figure 7: The synthesized images with "green" bounding box are with lower FRD scores, in which identity features are preserved better. However, FID and IS metrics are not aware of identity preserving.

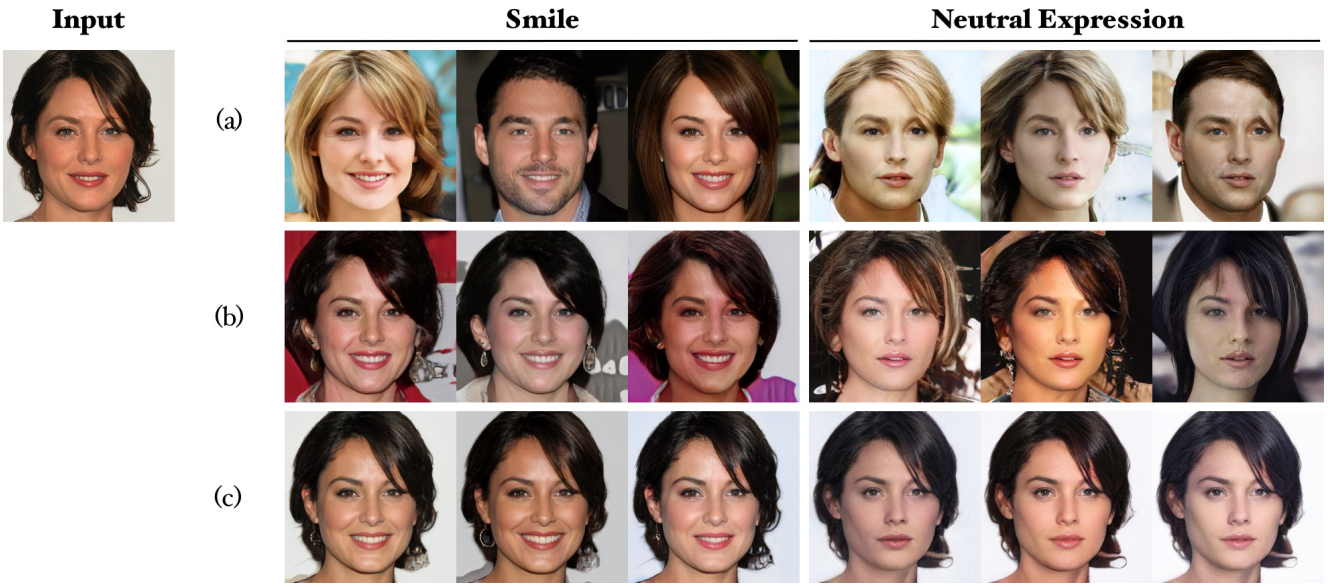


Figure 8: A visual comparison for the smile translation task on the CelebA-HQ dataset. (a) StarGAN v2 [3], (b) our proposed method without $\mathcal{L}_{cont}$ and (c) our proposed method with $\mathcal{L}_{cont}$. This comparison shows that a smooth style space can better preserve the person identity. Moreover, using $\mathcal{L}_{cont}$ significantly boosts the the input identity preservation.

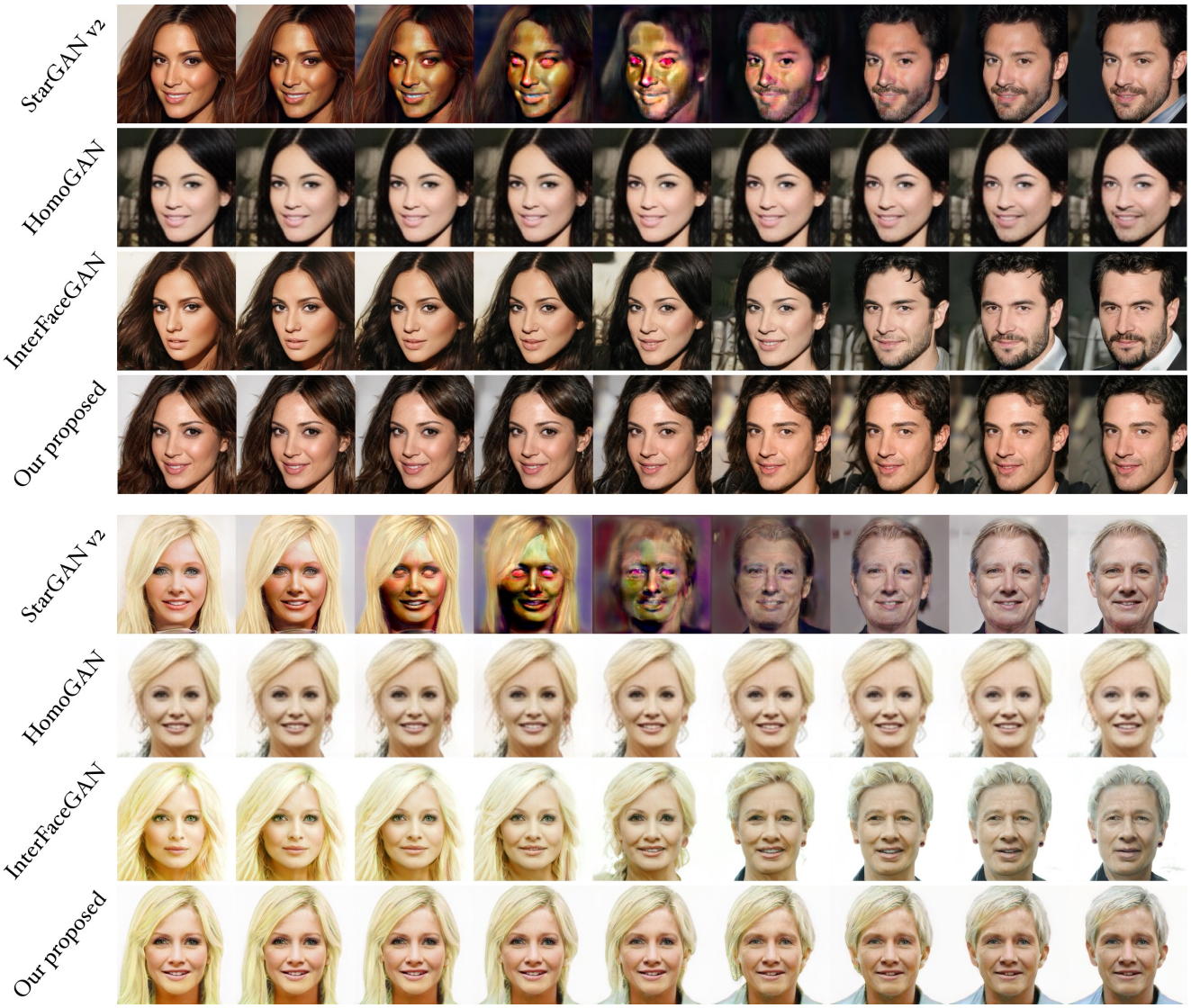


Figure 9: Additional comparisons between StarGAN v2 [3], HomoGAN [2], InterFaceGAN [10] and our proposed method on a gender translation task on the CelebA-HQ dataset [4].

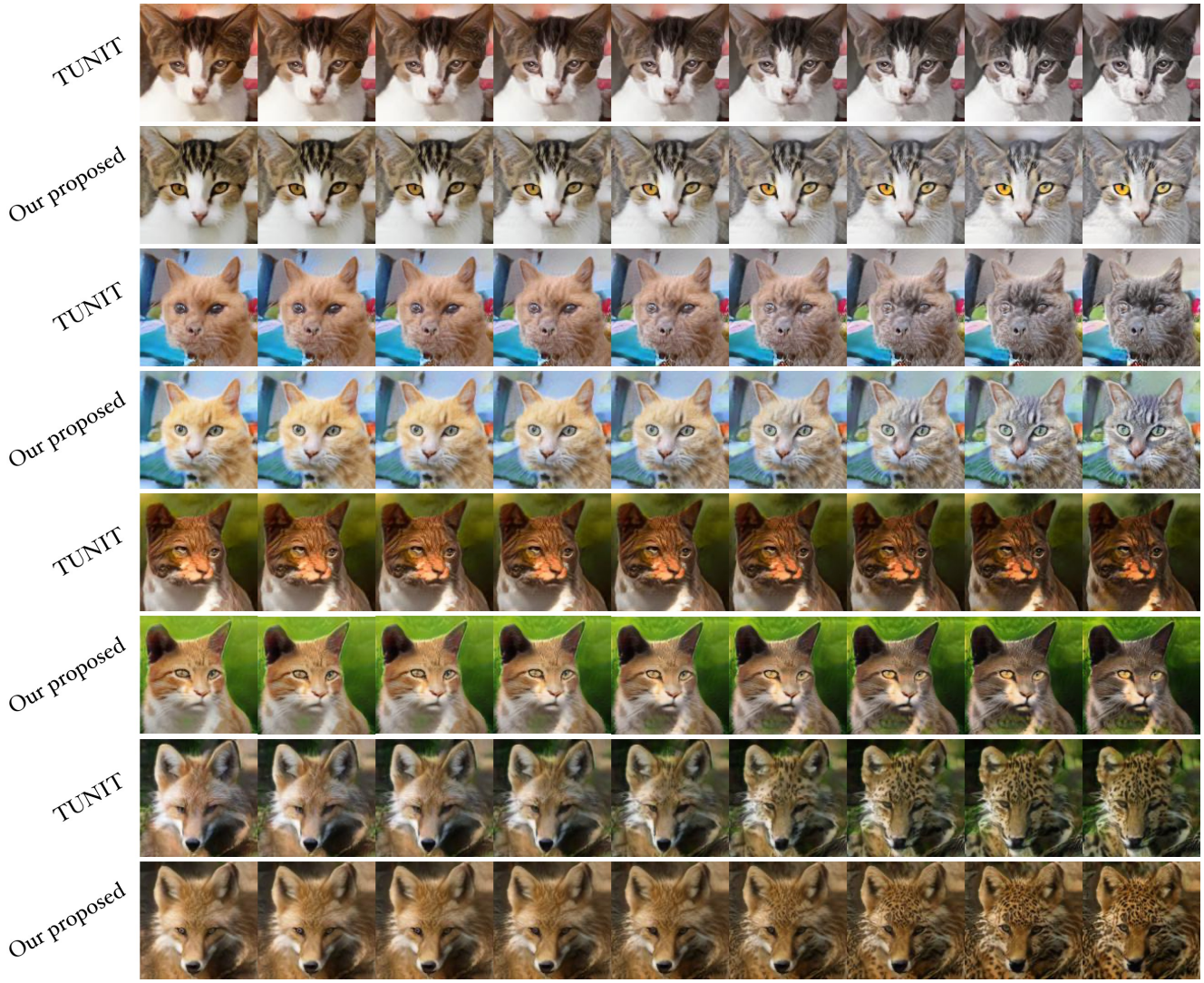


Figure 10: An additional comparison between TUNIT [1] and our proposed method on a truly unsupervised image-to-image translation task using the AFHQ dataset [3] (domain-level annotations are not provided).

Figure 11: More examples of gender translation on the CelebA-HQ dataset [4].

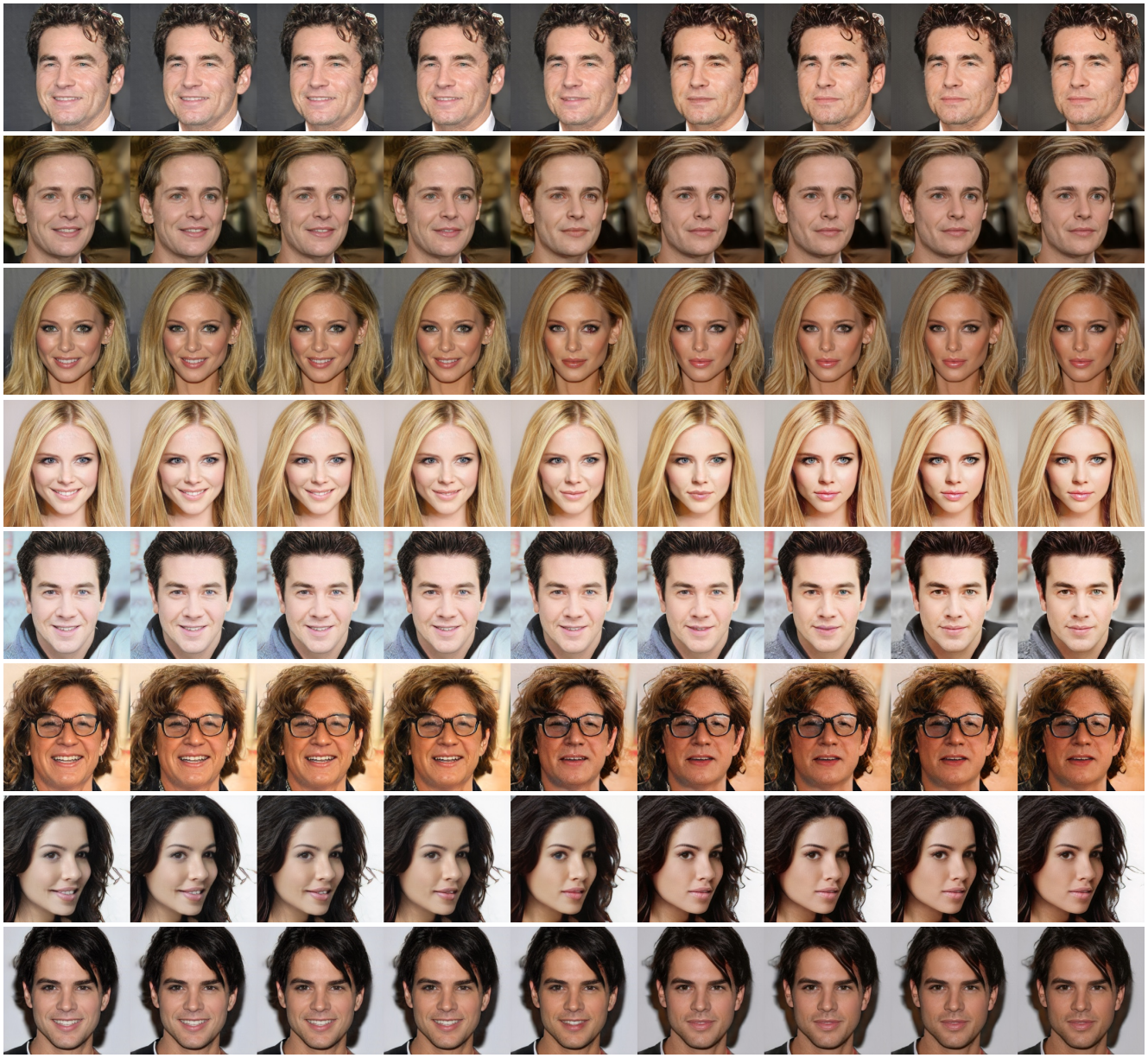


Figure 12: More examples of smile translations on the CelebA-HQ dataset [4].

Figure 13: More examples of age translations on the CelebA-HQ dataset [4].

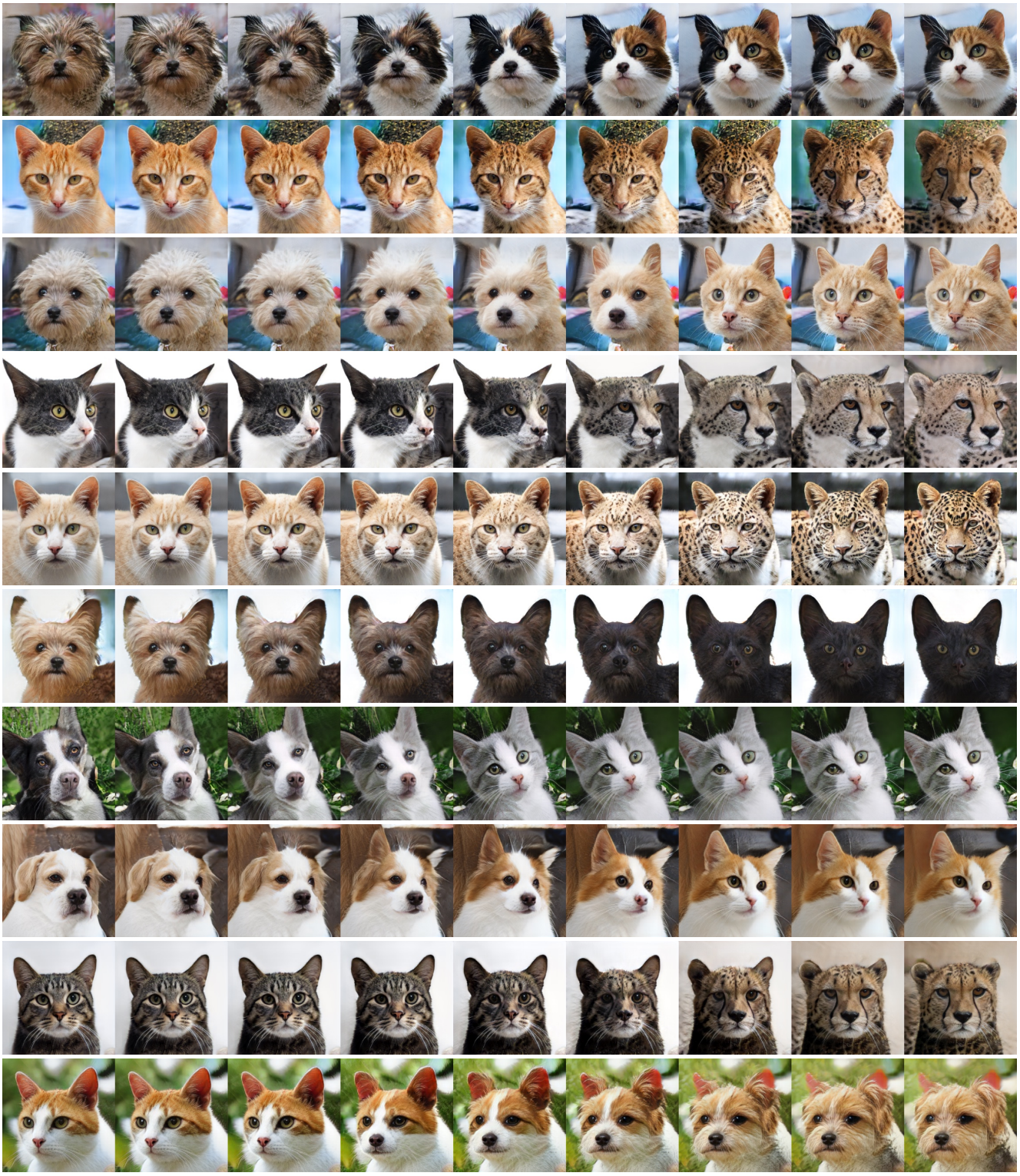


Figure 14: More examples of animal translations on the AFHQ dataset [3].

Figure 15: More examples of animal translations on the AFHQ dataset [3].

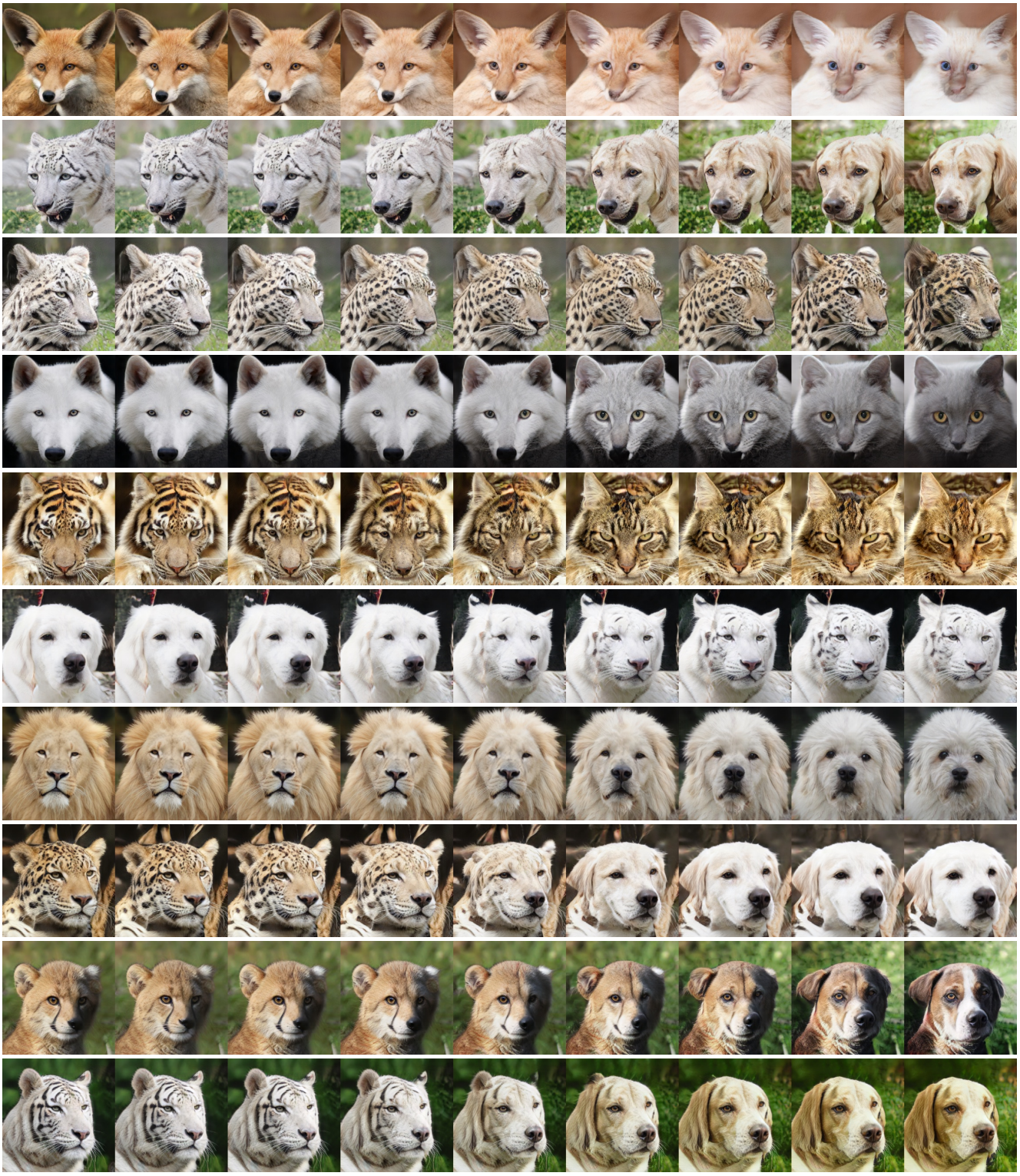


Figure 16: More examples of animal translations on the AFHQ dataset [3].

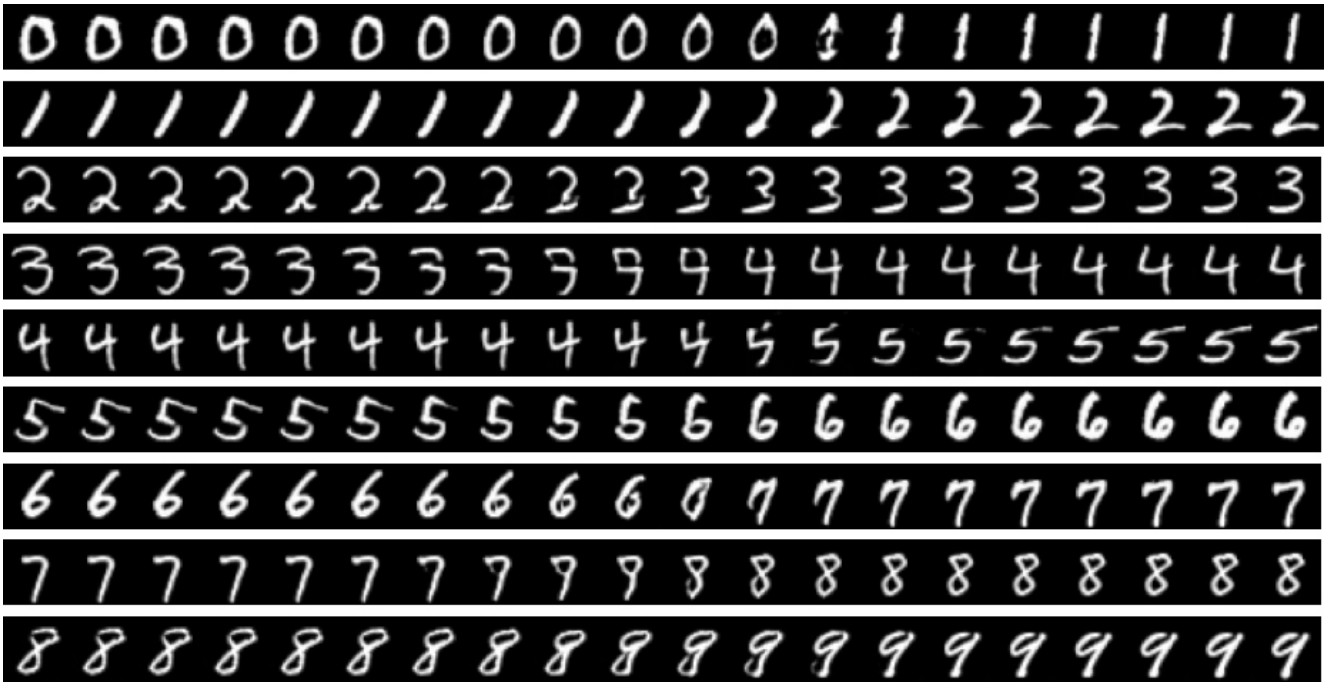


Figure 17: More examples of digits translations on the MNIST dataset [8].

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. *arXiv preprint arXiv:2006.06500*, 2020. 5, 8

[2] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019. 5, 7

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 3, 4, 5, 6, 7, 8, 12, 13, 14

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 5, 7, 9, 10, 11

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 4, 5

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[7] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5, 15

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[10] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 5, 7

[11] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3