One-Shot Generative Domain Adaptation
– *Supplementary Material* –

Ceyuan Yang[1,2,†] Yujun Shen[3,4,†] Zhiyi Zhang[3] Yinghao Xu[2] Jiapeng Zhu[5]
Zhirong Wu[6] Bolei Zhou[7]

[1]Shanghai AI Laboratory [2]CUHK [3]ByteDance Inc. [4]Ant Group [5]HKUST [6]MSRA [7]UCLA

A. Overview

This appendix is organized as follows: Sec. B clarifies the main differences from two relevant prior work, with the quantitative comparisons. Sec. C includes more implementation details. Sec. D conducts ablation studies on the proposed approach, including the *attribute adaptor*, the *attribute classifier*, the *diversity-constraint strategy*, and the hyper-parameter β in Eq. (5). Sec. E presents the quantitative comparison to prior approaches, demonstrating the effectiveness of our GenDA under the general few-shot setting. Sec. F visualizes the latent representations after adaptation. Sec. G analyzes the transformation (*i.e.*, the learnable vector $\mathbf{a}$ in Eq. (3) for attribute adaptation, which provides insights on how our algorithm works and shows how the latent representations are transformed after adaptation. Sec. H presents the additional comparison with the style transfer method and investigates the visual relation between and after adaptation.

Table A1: **Comparison against relevant baselines.**

One-shot (FID$_\downarrow$)	MineGAN	GenDA	FD	FD-all	FD-trunc.
Babies	145.21	80.16	147.96	156.82	132.13
Sunglasses	98.34	44.96	87.32	96.48	81.94
Sketches	109.85	87.55	102.13	113.61	98.59

Table A2: **Ablation study on architecture designs.** A two-layer multi-layer perceptron (MLP) is introduced as the heavy version of the attribute adaptor and the attribute classifier, respectively. FID (lower is better) is reported.

Models	**Babies**	**Sunglasses**	**Sketches**
Heavy attribute adaptor	290.22	262.32	114.25
Heavy attribute classifier	132.14	115.88	109.23
Lightweight design	**80.16**	**44.96**	**87.55**

Table A3: **Ablation study on attribute classifier of discriminator.**

One-shot on Babies	FID$_\downarrow$	Prec.$_\uparrow$	Recall$_\uparrow$
Fine-tuning original one	92.53	0.65	0.019
Attribute classifier (Ours)	80.16	0.74	0.033

Table A4: **Ablation study on the strength of diversity-constraint strategy**, which is noted as β in Eq. (5). Here, a smaller β denotes a stronger diversity constraint. Evaluation metrics include FID (lower is better), precision (higher means better quality), and recall (higher means higher diversity).

β	FID	Precision	Recall
0.5	24.87	0.82	0.163
0.7	22.62	0.78	0.204
0.9	24.32	0.73	0.269
1.0	25.29	0.71	0.349

B. Differences from Prior Efforts

This work tackles how to fine-tune a generator under a limited data regime (*i.e.*, as few as only one image). To alleviate the main problem (*i.e.*, the diversity collapse caused by over-fitting), we follow the commonly used strategy in prior literature, which is to reduce the number of learnable parameters. Under such a case, however, we manage to reuse the prior knowledge *to the most extent* (*i.e.*, only training *one layer* in the generator and the discriminator each), and confirm that such an efficient scheme indeed facilitates generative domain adaptation and substantially outperforms existing alternatives. Here, we further clarify the differences from two previous work that also consider to freezing parameters *i.e.,* FreezeD [8] and MineGAN [13].

Differences from MineGAN. 1) MineGAN *works in two stages*, where it first learns latent space transformation (with a frozen generator) and then fine-tunes the generator (with a frozen transformation). Differently, our approach trains the adaptor *once* with the generator frozen all the time. 2) Discriminator is *frozen* in our method but *fine-tuned* in MineGAN, which might lead to overfitting, especially under the one-shot setting. The first two columns of Tab. A1 present the one-shot comparison result where GenDA surpasses MineGAN in all three categories.

Differences from FreezeD. 1) In terms of the discriminator, FreezeD (FD) always keeps the *lower-level* features un-

touched and fine-tunes the *mid/high-level* features. Tab. A1 suggest that freezing all parameters ("FD-all") except the last layer would lead to a performance drop. 2) More importantly, our attribute classifier is *randomly initialized* which drops out the knowledge of original domain and pours more attention on the new domains. Tab. A3 demonstrates the superiority of our attribute classifier over directly fine-tuning the original last layer of discriminator. 3) Different from FreezeD, our GenDA also freezes generator and equips it with an attribute adaptor and diversity-truncation strategy, achieving appealing performances. In particular, the diversity-truncation strategy could also help the original FreezeD slightly, shown in the "FD-trunc." column of the above table.

C. Implementation Details

Implementation. We choose the state-of-the-art StyleGAN2 [4] as our base GAN model, following [10]. StyleGAN2 proposes a more disentangled $\mathcal{W}$ space in addition to the native latent space $\mathcal{Z}$. Here, our attribute adaptor is deployed on the $\mathcal{W}$ space. $\bar{\mathbf{z}}$ becomes $\mathbf{w}_{avg}$, which is a statistical average in the training of the source generator. To alleviate the overfitting problem, we apply differentiable augmentation on both the reference image and the adapted synthesis [17, 2]. In particular, we adopt the augmentation pipeline from StyleGAN2-ADA [2] and linearly increase the augmentation strength from 0 to 0.6. We use Adam optimizer [5] for training and the learning rates are set as $1.25e^{-4}$ and $2.5e^{-4}$ for $A(\cdot)$ and $\phi(\cdot)$, respectively. All experiments are stopped when the discriminator has seen $200K$ real samples. The last "ToRGB" layer (with 1×1 convolution kernels) of the generator is tuned together with the adaptor $A(\cdot)$ to make sure aligning the synthesis color space to the reference image. The hyper-parameter β in Eq. (5) is set as 0.7.

Datasets. Our source models are pre-trained on large-scale datasets (*e.g.*, FFHQ [3] and LSUN Church [16]) with resolution 256×256. We collect target images from FFHQ-babies, FFHQ-sunglasses, face sketches ([10]), and samples from masterpieces (*i.e.*, Mona Lisa and Van Gogh's houses) and Artistic-Faces dataset.

Evaluation Metrics. Fréchet Inception Distance (FID) [1] serves as the main metric. Akin to [10], we always calculate the FID between 5,000 generated images and all the validation sets. Additionally, we report the precision and recall metric [6] to measure the quality and diversity respectively.

D. Ablation Study

Component Ablation. In order to further investigate the role of each proposed component (*i.e.*, attribute adaptor, attribute classifier, and truncation strategy), we conduct comprehensive ablation studies on 1-shot and 10-shot adaptation. In particular, we choose FreezeD [8] as our baseline which freezes the lower features of the discriminator while fine-tunes the rest of the discriminator and the entire generator. Moreover, we also apply adaptive discriminator augmentation strategy (ADA) on it since data augmentations usually help alleviate the overfitting problem.

Table A5: **Ablation study on the components proposed in GenDA under one-shot**, including the attribute adaptor (AA), the attribute classifier (AC), and the diversity-constraint strategy (DC). Evaluation metrics include FID (lower is better), precision (higher means better quality), and recall (higher means higher diversity).

AA	AC	DC.	FID$_\downarrow$	Prec.$_\uparrow$	Recall$_\uparrow$
			147.91	0.61	0.012
✓			92.54	0.53	0.162
✓	✓		86.21	0.63	**0.176**
✓	✓	✓	**80.16**	**0.74**	0.033

Table A6: **Ablation study on the components proposed in GenDA under 10-shot**, including the attribute adaptor (AA), the attribute classifier (AC), and the diversity-constraint strategy (DC). Evaluation metrics include FID (lower is better), precision (higher means better quality), and recall (higher means higher diversity).

AA	AC	DC.	FID$_\downarrow$	Prec.$_\uparrow$	Recall$_\uparrow$
			109.87	**0.80**	0.000
✓			53.59	0.57	0.064
✓	✓		50.37	0.62	**0.106**
✓	✓	✓	**47.05**	0.71	0.065

As is shown in Tab. A5 and Tab. A6, the baseline alternative could achieve the satisfying synthesis quality but worse diversity, suggesting that it might overfit the training shot significantly. By involving the attribute adaptor, the baseline is substantially improved, which already outperforms prior approaches. Meanwhile, the boosted diversity also demonstrates our motivation that we could reuse the variation factors of the source model and hence maintain the synthesis diversity to some extent. However, the quality (*i.e.*, precision) on babies and sketches becomes worse. After being equipped with the attribute classifier, the overall measurement is further enhanced. Together with the diversity-constraint strategy, our GenDA could facilitate the quality to the baseline level but achieve better diversity, leading to the new state-of-the-art few-shot adaptation performances.

Architectural Ablation. To further study the effect of the lightweight design of the attribute adaptor and the attribute classifier, we conduct experiments of replacing the lightweight module with a two-layer multilayer perceptron (MLP), which has a higher learning capacity. Tab. A2 presents the results. Obviously, with either a heavier

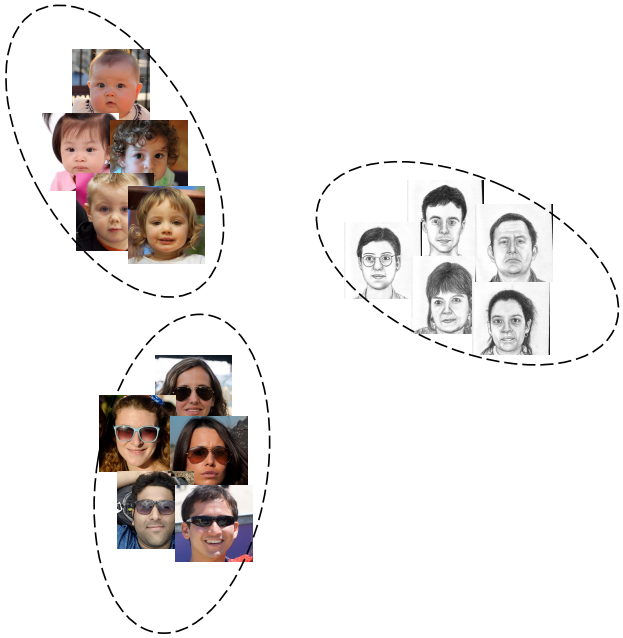

Figure A1: **Analysis on the transformation vector for domain adaptation**, which is noted as $\mathbf{a}$ in Eq. (3). We launch 15 one-shot adaptation experiments with 5 kids, 5 sunglasses, and 5 sketches as the target, respectively. After the model converges, we collect the transformation vector learned by the *attribute adaptor*, and then perform PCA on all 15 vectors. We can tell that our GenDA tends to learn similar transformation regarding the reference with similar characters. The 15 training samples are visualized together with the PCA results.

attribute adaptor or a heavier attribute classifier, the synthesis performance drops significantly. This might be caused by the overfitting problem since we only have one training image under our one-shot setting. It verifies the effectiveness and necessity of our lightweight design for the challenging one-shot generative domain adaptation task.

Hyper-parameter Ablation. We also conduct the ablation of diversity constraint parameter β in Eq. (5). Specifically, we train our model with 10-shot sunglasses images and maintain the evaluation metrics used in Sec.3.2. As is shown in Tab. A4, when increasing the value of β, the synthesis quality becomes worse while the diversity is enhanced. Noticeably, when the β is less than 0.5, the training diverges. Therefore, we choose 0.7 in all experiments for its overall performance.

E. 10-shot Domain Adaptation

As our GenDA could capture the representative attributes precisely with an increasing number of target shots, we also compare again prior approaches under the general few-shot adaptation setting. Following the setting of [10], we choose 10-shot target domain for comparison.

Table A7: **Quantitative comparison on 10-shot adaptation.** FID (lower is better) serves as the evaluation metric

10-shot transfer	**Babies**	**Sunglasses**	**Sketches**
TGAN [14]	104.79	55.61	53.41
TGAN+ADA [2]	102.58	53.64	66.99
BSA [9]	140.34	76.12	69.32
FreezeD [8]	110.92	51.29	46.54
MineGAN [13]	98.23	68.91	64.34
EWC [7]	87.41	59.73	71.25
Cross-Domain [10]	74.39	42.13	45.67
GenDA (ours)	**47.05**	**22.62**	**31.97**

Quantitative Comparison. Multiple baselines are introduced to conduct the quantitative comparisons, including Transferring GANs (TGAN) [14], Batch Statistics Adaptation (BSA) [9], FreezeD [8], MineGAN [13], EWC [7] and Cross-Domain [10]. All these methods adapt a pre-trained source model to a target domain (*e.g.*, babies, sunglasses and sketches). Akin to Cross-domain [10], FID serves as the metric for evaluation.

Tab. A7 presents the 10-shot quantitative comparison on several target domains. Although Cross-Domain [10] has already obtained the competitive results, our method could still significantly improve the performance by a clear margin, resulting in the new state-of-the-art synthesis quality on few-shot adaptation. Specifically, for domains that differ in semantic attributes (*i.e.*, babies and sunglasses), huge gains are obtained by GenDA. It is also worth noting that when the domain gap increases (*i.e.*, from realistic facial images to sketches), the performances of Cross-domain [10] and FreezeD [8] become almost identical (45.67 *v.s* 46.54) while substantial improvements of synthesis could remain observed by our method (31.97). It demonstrates the effectiveness of our method on the general few-shot settings.

F. Properties of Adapted Latent Spaces

In order to leverage the generative models for content editing, prior work [11, 15] proposed to find the semantic directions in the latent spaces of a well-trained generative model. Such that, we could adjust a certain attribute of the synthesis by moving its corresponding latent code in one of these semantic directions. As our motivation is to reuse the knowledge of source models, we therefore investigate whether the latent space after adaptation remains to maintain the same semantic directions.

We start with latent space interpolation which can reflect whether the generative model overfits the training samples. In particular, all models are fine-tuned with one shot image. Fig. A4 presents the generated results during interpolation between two latent codes. The changes are smooth and

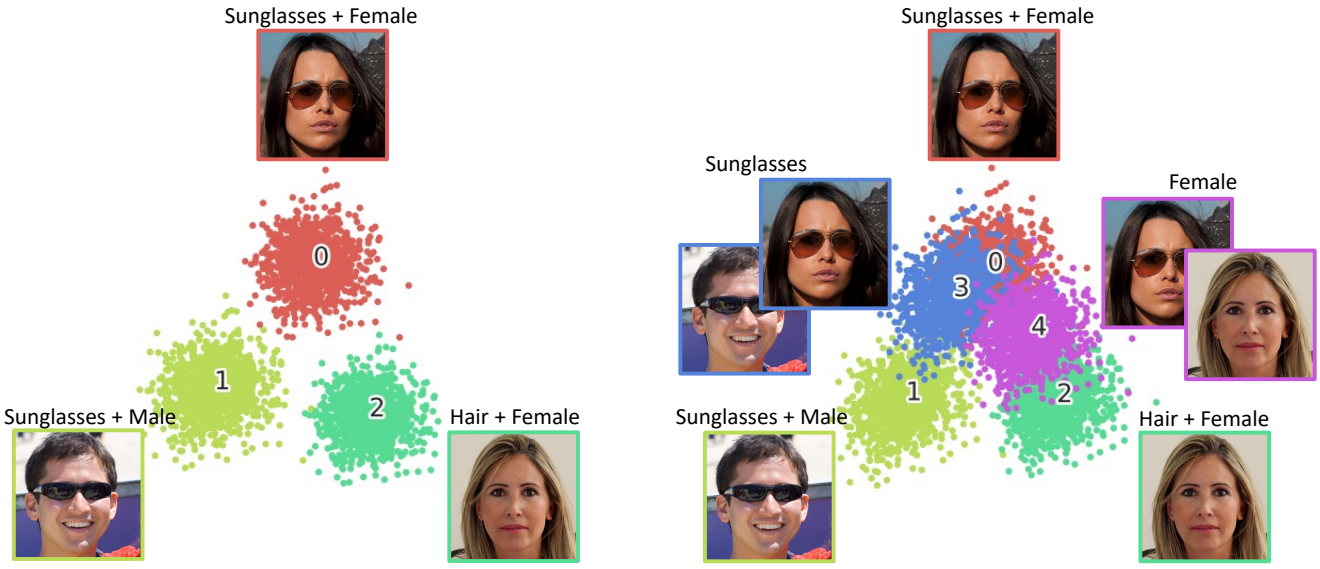

Figure A2: **Visualization of adapted latent spaces after domain adaptation.** Each cluster corresponds to a particular training set. Left: The latent spaces learned with different reference images are clearly separable. Right: When we use two reference images, which have common attributes, as a combined training set, the transformed representations tend to locate at the overlapped region between the representations learned from each reference image independently.

Figure A3: **Latent representation comparison before and after adaptation.** The latent space is clearly pushed from the **red** cluster to the **blue** cluster when there exists an obvious gap between the source and the target domains.

Figure A4: **Properties of adapted latent spaces.** Note that, all models (except the source model) are trained with one shot image. Top: during the interpolation between two latent codes (corresponding to left and right most images), the resulting images change smoothly and within target domain. Bottom: GenDA preserves the semantic directions of source generative models which are applicable on models after adaptation

within target domains (from top to bottom: sunglasses, babies, and sketches), indicating the attributes are well-captured by our adaptation.

Then, we investigate the semantic properties of the adapted generative models. We first present the effectiveness of semantic controlling via InterFaceGAN [11] on the source model. As shown in the middle of Fig. A4, the semantic direction of smiling control works well. If

we directly apply this direction on three different target models, they all reveal the same semantic changes while remaining other attributes (shown in the bottom of Fig. A4). We accredit the preservation of semantic direction to the design of inheriting knowledge from source domain and the lightweight modules of GenDA. Such a property can be of great benefit to content manipulation on various target domains, with the semantic discovery performed on source domain for only once.

G. Analysis

Transformation Vector. We also visualize the learned attributes via Principal components analysis (PCA). Concretely, we first choose 15 shots from 3 domains (*i.e.*, babies, sunglasses and sketches). The corresponding models are trained on them and 15 attribute vectors in Eq. (3) are collected to perform PCA. Fig. A1 presents the PCA results. Obviously, the attribute vectors of different shots but the same domain assemble closely, while there are obvious decision boundaries across domains. Such interpretation happens to match our motivation that the attribute adaptor could capture the representative attributes and make the corresponding adjustment for a source model to multiple target domains.

Transformed Latent Representation. We also visualize how the latent representations change before and after adaptation, with the help of PCA. Specifically, we sample $2K$ latent codes from the latent space after adaptation for each model used in Fig.4 and Fig.5. Fig. A2 suggests that the distributions after adapting the source model to three different target domains are clearly separable. Also, when using more than one reference image, which have common attributes, from the target domain, the transformed representations tend to locate at the overlapped region between the representations learned from each reference image independently. Our supplementary material also presents the latent representation change after cross-domain adaptation, where the latent distribution of the source domain is successfully shifted to the target one.

We also visualize the latent representation change for cross-domain adaptation. Fig. A3 shows the latent representation change after cross-domain adaptation, where the latent distribution of the source domain is successfully shifted to the target one.

H. Additional Results

Comparison with the style transfer task. As discussed in Sec.3.3, the proposed approach might be an alternative for style transfer. Here, we qualitatively compare our approach with the state-of-the-art method, SWAG [12] regarding the style transfer task. The comparison results are included in Fig. A5. From the perspective of the color transfer, SWAG [12] does better since it could successfully transfer the representative color from the target image to the synthesis. For example, the green of the vegetation and the pink of the skin appear in the transferred samples. However, it also reveals its weakness that the transfer is inconsistent. Specifically, the wall or the sky in church synthesis is in pink while our method pictures them in the same color. This implies that the style-transfer-based methods could have no semantic concepts regarding the content, leading to such inconsistency and obvious artifacts. Therefore, it naturally cannot transfer higher-level attributes like sunglasses while ours could.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[2] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Adv. Neural Inform. Process. Syst.*, 2020. 2, 3

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[6] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *arXiv preprint arXiv:1904.06991*, 2019. 2

[7] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[8] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 1, 2, 3

[9] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Int. Conf. Comput. Vis.*, 2019. 3

[10] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3

[11] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 3, 5

[12] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In

Figure A5: **Comparison with the style transfer task.** GenDA could transfer the characteristic of the target image to the source model in a more harmonious way, while SWAG [12] transfers the color inconsistently, leading to obvious artifacts.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 124–133, 2021. 6, 7

[13] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 3

[14] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Eur. Conf. Comput. Vis.*, 2018. 3

[15] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *Int. J. Comput. Vis.*, 2021. 3

[16] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans

in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2

[17] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Adv. Neural Inform. Process. Syst.*, 2020. 2