TEXT-DRIVEN GENERATIVE DOMAIN ADAPTATION WITH SPECTRAL CONSISTENCY REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Combined with the generative prior of pre-trained models and the flexibility of text, text-driven generative domain adaptation can generate images from a wide range of target domains. However, current methods still suffer from overfitting and the mode collapse problem. In this paper, we analyze the mode collapse from the geometric point of view and reveal its relationship to the Hessian matrix of generator. To alleviate it, we propose the spectral consistency regularization to preserve the diversity of source domain without restricting the semantic adaptation to target domain. We also design granularity adaptive regularization to flexibly control the balance between diversity and stylization for target model. We conduct experiments for broad target domains compared with state-of-the-art methods and extensive ablation studies. The experiments demonstrate the effectiveness of our method to preserve the diversity of source domain and generate high fidelity target images.

Figure 1: Text-driven generative domain adaptation with various text descriptions. The generated samples both reflect characteristics of target domain from text and preserve the original identity.

1 INTRODUCTION

Generative image modeling has developed significantly in recent years and is able to generate diverse high-resolution images even indistinguishable from real images. However, training such models requires intense computation resources and large datasets, which restricts the application scope of generative models. For some scenarios, collecting large datasets is impossible like paintings by specific artists. Benefiting from Vision-Language models learning from large image-text pairs, text can be leveraged as a description of abstract visual semantics to guide generative domain adaptation instead of a collection of image samples. As an expressive representation, text has shown great success in semantic image generation and manipulation recently (Saharia et al., 2022; Ramesh et al., 2022). Based on the generative prior of pre-trained models and flexible text description of target domain, text-driven domain adaptation can generate more various images and have promising applications.

To reduce the requirement of training samples, traditional methods propose to train generative models in the target domain with only limited samples by adapting pre-trained models in the large-scale source domain which contains high-level semantic knowledge as a generative prior. These few-shot adaptation methods either finetune only a part of parameters within networks to preserve most source domain knowledge (Mo et al., 2020) or impose strong regularization on the generated images (Xiao et al., 2022; Zhu et al., 2021). However, these methods still require additional training samples

of target domain and adversarial training process. As the number of samples drops, the image fidelity and diversity also hurt severely. Different from these methods, text-driven domain adaptation requires no image samples but texts to describe the target domain. Pioneer work (Gal et al., 2021) proposed to encourage the visual change between samples from target and source generators to align with semantic direction described by text in the CLIP (Radford et al., 2021) embedding space, which achieves generative adaptation for miscellaneous domains in short training time.

The main challenge of text-driven GAN adaptation is the mode collapse problem due to the entanglement of intra-domain semantics and inter-domain style in text representation. Besides the specified target style described by text, there also exists an unknown pattern for the semantics of images. This leads to a decrease of variations in generated images when the style effect is optimized to approach target domain. As shown in Figure 2, while the number of iterations increases, the generated sample tends to have similar patterns of mouth and eyes, which reduces most of the variations in the origin model. The main reason for the mode collapse problem is that the optimization process only cares about the distance of generated samples to target domains, and the intra-domain feature variations are easily ignored.

To address the above challenge, Zhu et al. (2021) proposed to preserve the diversity of source domain through a within domain loss which keeps consistency between sample changes in source domain and target domain. However, this regularization is too strong to restrict the style effect of target generator close to source domain. The previous theorem about GAN latents analysis has shown that the Hessian matrix of generator reflects the variations of generator and can be used to explore meaningful directions from top eigenvectors. Inspired by this, we try to leverage the spectrum of Hessian matrix as a quantitative evaluation of model diversity in the adaptation problem. This disentangles the relative diversity between generated samples from absolute generative distribution and makes a general way to regularize diversity of generative model.

In this work, we propose spectral consistency regularization to solve the problem of mode collapse in text-driven domain adaptation from the geometric point of view. First, we analyze the Hessian matrix of generator's manifold in the metric space by eigendecomposition. The eigenvalues of Hessian matrix are decreasing in the adaptation process, which is consistent with the mode collapse problem of visual observations. Second, we introduce the spectral consistency regularization on the Hessian matrix to prevent the latent space of generator from degrading. This regularization helps preserve intra-domain variations of source domain without restricting style effects of target generators. We further develop a stochastic method to regularize the spectrum of Hessian matrix without calculating the full matrix, which reduces the expensive computational cost. Finally, we design the granularity adaptive regularization considering the layer-decomposition characteristic of W+ space in StyleGAN.

In summary, our contributions of this paper are as follows:

1. We analyze the commonly occurred mode collapse problem in GAN adaptation from the geometric point of view and provide a quantitative evaluation of model diversity to reveal the reason of mode collapse.
2. We propose the spectral consistency regularization for text-driven generative domain adaptation, which both preserve the diversity of original domain and generates high fidelity images of target domain. A granularity adaptive regularization is further designed to flexibly control the balance between diversity and stylization for target model.
3. We conduct experiments and ablation studies for a wide range of target domains. The experiments show the effectiveness of our proposed spectral consistency regularization and its applications to downstream tasks like image editing and image-to-image translation.

2 Related Work

Text-driven Image Synthesis and Manipulation Traditional methods approached text-driven image generation by training a conditional GAN(Reed et al., 2016). Several following works have been proposed to improve generation quality either by multi-scale networks (Zhang et al., 2017) or attention mechanism (Xu et al., 2018). Recently, transformer-based auto-regressive generative models were introduced to view text-driven image synthesis as conditional sequence generation of visual tokens conditioning on text embeddings (Esser et al., 2021; Ramesh et al., 2021; Yu et al.,

2022). Diffusion models were also leveraged as the decoder for image generation, which achieves tremendous improvement for generating high quality images (Saharia et al., 2022; Ramesh et al., 2022).

Another kind of method is to leverage Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) models as knowledge guidance for text based image generation. This is achieved by optimizing the latent code of pretrained generator to bring close the distance between generated images and input text in the shared embedding space. The optimized latent codes of generator are either in the StyleGAN latent space (Patashnik et al., 2021) or VQGAN codebook space (Crowson et al., 2022). Some methods also defined the parametrized space by vector graphics such as Bezier curves (Frans et al., 2021).

Few-shot GAN Adaptation aims to transfer pretrained generator to another domain when there are not enough samples to train from scratch. Its main challenge is the mode collapse problem because the generator is prone to overfit training samples in target domain and lose the diversity of the origin domain. There have been many methods to tackle this problem. They either froze most of the parameters of the pre-trained network(Mo et al., 2020) or embedded a small number of trainable parameters into the source model(Noguchi & Harada, 2019). Recently, Ojha et al. (2021) proposed cross-domain distance consistency loss to preserve the relative similarities and differences between instances in the source domain. Xiao et al. (2022) introduced spatial structural consistency loss to align the spatial information between the synthesis image pairs of the source and target domains. These methods still require manually collected samples, and as the number of samples decreases, the mode collapse problem becomes more apparent.

Besides, StyleGAN-NADA (Gal et al., 2021) further proposed to take advantage of the CLIP (Radford et al., 2021) model as knowledge guidance for GAN adaptation, and only natural language prompts are required without even a single image. Similarly, Zhu et al. (2021) used the image encoder of CLIP for one-shot adaptation. However, these methods still suffer from the mode collapse problem. In this paper, we propose the spectral consistency regularization to tackle the problem of mode collapse without hurting target generation performance.

Latent Space Analysis of GANs Many works have explored the latent space of pretrained generator for image manipulation. Some methods used supervised datasets to learn directions in the latent space for attribute editing (Shen et al., 2020) or semantic image editing (Ling et al., 2021). Other works instead applied unsupervised methods to reveal the latent space. Shen & Zhou (2021) decomposed the learned weights of the pre-trained network to identify semantically meaningful directions. Härkönen et al. (2020) applied principal component analysis in the latent space. Recently, Wang & Ponce (2021) proposed to analyze the latent space of generative models from geometric point of view. They found that the eigenvectors corresponding to the largest eigenvalues of the Hessian matrix for generator dominate interpretable variations. In this paper, we analyze the GAN adaptation problem in a similar way and propose to regularize target generator by the spectrum of Hessian matrix.

3 METHOD

3.1 TEXT-DRIVEN GENERATIVE DOMAIN ADAPTATION

Text-driven domain adaptation aims to transfer a pretrained generator to target domain specified by the text description. To guide the domain adaptation by text, pre-trained CLIP model is leveraged to measure the similarity between image and text. CLIP is a Vision-Language model trained on 400 million (image, text) pairs collected from the internet with contrastive loss (Radford et al., 2021). One commonly used objective function for text-driven image manipulation is the global loss that optimizes the similarity between generated images and target text:

$$\mathcal{L}_{global} = D_{CLIP}(G(z), t_{target}) \tag{1}$$

where D_{CLIP} is the cosine distance in the CLIP space, t_{target} is the target text. However, this only applies to in-domain image manipulation combined with identity consistency regularization(Patashnik et al., 2021), and this regularization is too strong for cross-domain adaptation with large domain gaps like human to werewolf. Direct optimization of the above global loss leads to adversarial solutions since adding pixel-level perturbations can fool the CLIP classifier in the absence

of a generative prior favoring real-image manifold (Gal et al., 2021). To overcome this limit, the directional loss is used to optimize the direction between source and target domain:

$$\mathcal{L}_{\text{direction}} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|}$$
$$\Delta T = E_T\left(t_{\text{target}}\right) - E_T\left(t_{\text{source}}\right), \Delta I = E_I\left(G_{\text{train}}\left(z\right)\right) - E_I\left(G_{\text{frozen}}\left(z\right)\right)$$

where E_T and E_I are the text and image encoder of CLIP, t_{source} and t_{target} are the source and target class texts. The directional loss can prevent adversarial solutions. Since target generated images should have given direction to corresponding source images, generating a single adversarial instance is impossible. At training time, the same latent code is fed into source and target generator, then the target generator is optimized using the directional loss. However, the directional loss proposed by StyleGAN-NADA still suffers from the mode collapse problem, as shown in the images of Figure 2. So we propose the spectral consistency regularization derived from geometry analysis of GAN to resolve this problem.

3.2 Geometry Analysis of GAN Adaptation

We denote the generative network as a mapping from latent code z to a manifold in image space as $G(z)$. Considering a squared distance function d^2 for two images, we express the local variations of $G(z)$ from moving towards direction Δz by second-order Taylor expansion. This is formulated as:

$$\lim_{\Delta z \to 0} d^2(G(z), G(z+\Delta z)) = d^2(G(z), G(z)) + \frac{\partial d^2(G(z), G(z+\Delta z))}{\partial \Delta z} \cdot \Delta z + \Delta z^T \cdot \frac{\partial^2 d^2(G(z), G(z+\Delta z))}{\partial \Delta z^2} \cdot \Delta z \tag{2}$$

The first two terms are zero since $d^2(G(z), G(z+\Delta z))$ is local minima when $\Delta z = 0$. Denote the second derivatives as Hessian matrix $H(z)$, and we have $d^2(G(z), G(z+\Delta z)) = \Delta z^T H(z) \Delta z$. For a normalized vector Δz, we can conclude that $\sigma_{min} \leq d^2(G(z), G(z+\Delta z)) \leq \sigma_{max}$, where σ_{min} and σ_{max} are the smallest and largest eigenvalues of H. Thus, we can use the trace norm of H_z to reflect the statistics of diversity in generative models, which is the sum of all eigenvalues of H. Especially, for a squared $L2$ distance function in metric space ϕ, $d^2(z_1, z_2)_\phi = \frac{1}{2}\|\phi(G(z_1)) - \phi(G(z_2))\|^2$, the Hessian matrix $H_\phi(z_0)$ is a simple transformation from the Jacobian $J_\phi(z_0)$:

$$H_\phi\left(\boldsymbol{z}_0\right) = \frac{\partial^2}{\partial \boldsymbol{z}^2} \frac{1}{2} \left\|\phi\left(\boldsymbol{z}_0\right) - \phi(\boldsymbol{z})\right\|_2^2 = J_\phi\left(\boldsymbol{z}_0\right)^T J_\phi\left(\boldsymbol{z}_0\right), \tag{3}$$

$$\boldsymbol{v}^T H_\phi\left(\boldsymbol{z}_0\right) \boldsymbol{v} = \left\|J_\phi\left(\boldsymbol{z}_0\right) \boldsymbol{v}\right\|^2, \quad J_\phi\left(\boldsymbol{z}\right) = \frac{\partial \phi(G(z))}{\partial z}, \tag{4}$$

and the top eigenvectors of H_ϕ correspond to right singular vectors of the Jacobian J_ϕ.

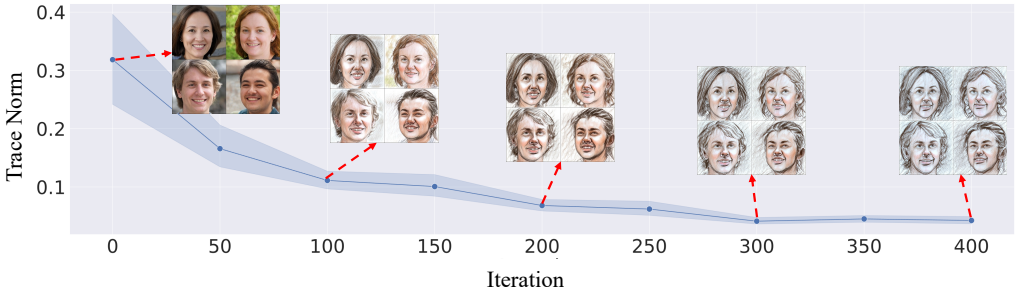

Figure 2: The mode collapse problem in the photo-to-sketch domain adaptation by StyleGAN-NADA Gal et al. (2021). The trace norm of Hessian matrix is gradually decreasing, which is consistent with the visual examples showing similar patterns with mouth and eyes.

To analyze the mode collapse problem, we calculate the statistics of Hessian trace from different samples of z during adaptation process of the previous state-of-the-art method. The results of domain adaptation from photo to sketch with StyleGAN-NADA Gal et al. (2021) is shown in Figure 2. We find that as the iteration count increases, the mode collapse problem becomes more severe and the Hessian trace is also decreasing, which proves that the Hessian trace can reflect the diversity of generator. In the early stage of adaptation, the decrease of trace norm are mainly caused by style adaptation since there are no color variations for sketch domain. But for the late stage from step 200 to step 400, the style effect changes little and the structures tend to have the same pattern. Since the target text only represents a fixed direction without variation, different samples are encouraged to approach the same fixed pattern.

3.3 Spectral Consistency Regularization

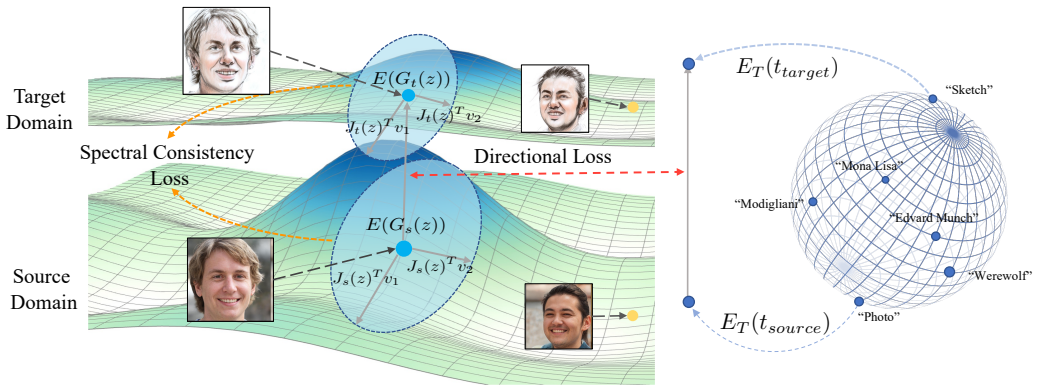

Figure 3: Illustration of our proposed method for text-driven generative domain adaptation. The training objective of our model consists of directional loss and spectral consistency loss. We feed the same latent code to source generator and target generator and generates a pair of source and target images. The directional loss encourages the direction between embeddings of the pair to align with the semantic direction of text description. The spectral consistency loss regularizes the trace norm of Hessian matrix of target generator to prevent the mode collapse problem.

To prevent the target generator from mode collapse, we propose spectral consistency regularization to prevent the diversity of generator from degrading, which is calculated as:

$$L_{reg} = \|Trace(H_s(z)) - Trace(H_t(z))\|, \tag{5}$$

where $H_s(z)$ and $H_t(z)$ are the Hessian matrix of the source generator and target generator evaluated with the same latent code z. However, directly computing the Hessian matrix requires backpropagation n times where n is the dimension of feature vector in metric space. Instead, we use the Hutchinson's method for trace estimator (Hutchinson, 1989) to compute a stochastic estimator of Hessian Trace, which is formulated as:

$$Trace(H(z)) = \mathbb{E}[v^T H(z) v] = \mathbb{E}[\|J_\phi(z)v\|^2], \;\; v \sim \text{Rademacher}(\frac{1}{2}) \tag{6}$$

The second transformation is derived from Equation 2 and 3. So the calculation of Hessian matrix is transformed into the calculation of Jacobian-Vector product.

Different from the within domain loss(Zhu et al., 2021) which restricts target generated samples based on relative difference of source samples, the spectral consistency regularization only cares about the diversity of target model. This doesn't impose restriction on the direction of target adaptation, so our method can generate samples more consistent with target text without losing model diversity.

The training objective of text-driven generative domain adaptation is a weighted combination of directional loss and spectral consistency regularization loss $\mathcal{L} = \mathcal{L}_{dir} + \lambda \mathcal{L}_{reg}$. Since different target domains have their own characteristic, it is required to tune the hyperparameter λ for better

performance. To prevent the exhausting hyperparameter searching, we propose an adaptive loss reweighting method to balance the influence of these two loss items. Specifically, the adaptive weight λ is calculated as $\lambda_{spectral}\frac{\|\nabla_{G_L}\mathcal{L}_{dir}\|}{\|\nabla_{G_L}\mathcal{L}_{reg}\|}$, where G_L denotes the last layer of generator and $\lambda_{spectral}$ is a manually specified hyperparameter, typically 1.0 is an appropriate choice.

3.4 Granularity Adaptive Regularization

For text-driven domain adaptation problem, the adaptation granularity of different target domains varies from texture to structure. For example, the photo-to-sketch adaptation mainly focuses on the appearance and texture change, while the werewolf domain has more variations on semantic structure. Regularization of diversity in the whole granularity will restrict the adaptation performance. To alleviate this problem, we propose granularity adaptive regularization based on the disentangled characteristic of StyleGAN latent code.

The latent codes injected into different layers in StyleGAN influence different granularities, where the style code in low resolution represents high-level aspects such as pose and face shape, that in middle resolution controls facial features and hairstyle, and that in high resolution influences color scheme and microstructure. Specifically, we use the $W+$ space as input space for Jacobian matrix calculation, where each input latent code consists of 18 512-dimensional vectors so both z and v in Equation 4 are in $R^{18\times512}$. By masking v with a mask vector $m \in \{0,1\}^{18}$ for different layers , we can specify the granularity of variations involved in the regularization. The calculation of Hessian trace under mask is formulated as:

$$Trace(H(z)) = \mathbb{E}[(v\odot m)^T H(z)(v\odot m)] = \mathbb{E}[\|J_\phi(z)(v\odot m)\|^2],\ \ v\sim \text{Rademacher}(\frac{1}{2}). \quad (7)$$

Following previous convention, we divide the style code for 18 layers into 3 groups, which are for coarse, middle and fine scale. The results of different mask strategies are shown in Figure 6.

To explore the best mask strategy for different target domains, we propose to use an adaptive soft mask vector $\{\tilde{m}|\tilde{m}\in R^{18}, \|\tilde{m}\|=1\}$ for all layers. During training, the mask vector is optimized with respect to the overall training objective. To reduce the directional loss, the mask vector will assign less value to the latent code corresponding to the granularity that changes most, while other values will increase to preserve the diversity of source generator.

4 Experiment

In this section, we will show the qualitative and quantitative results of our method. We illustrate the generated results for a wide range of target domains from style and texture changes to shape and semantic modifications. We also compare the proposed spectral consistency regularization with other regularization methods for diversity preservation. Next, we perform an ablation study on our method to evaluate the effectiveness of each component. Finally, we demonstrate the applications of text-guidance domain adaptation, including image-to-image translation and image editing. The training details are explained in Appendix A.1. The choice of input space is explained in Appendix A.2.

4.1 Comparison Results

Qualitative Comparison In Figure 4, we show the comparison results of our method with state-of-the-art model StyleGAN-NADA Gal et al. (2021) for a wide range of target domains, which varies from texture changes like sketch and Mona Lisa paintings to geometric change like werewolf and Pixar style. The results demonstrate that our method not only generates highly stylistic images consistent with target text description for different target domains, but also produces images with diversity inherited from the pretrained source generator. Compared with StyleGAN-NADA which has obvious mode collapse problem like the mouth pattern in sketch and hair in werewolf, our model generates target images with better identity consistency. This proves that the spectral consistency regularization can preserve the diversity of source domain. In the Figure 9 and 10 of Appendix A.4, we present additional visual results for the dogs and cars domain.

We also perform domain adaptation experiments with other regularization methods, including the Selective Cross-modal Consistency (SCC) loss(Zhang et al., 2022), Within domain consistency

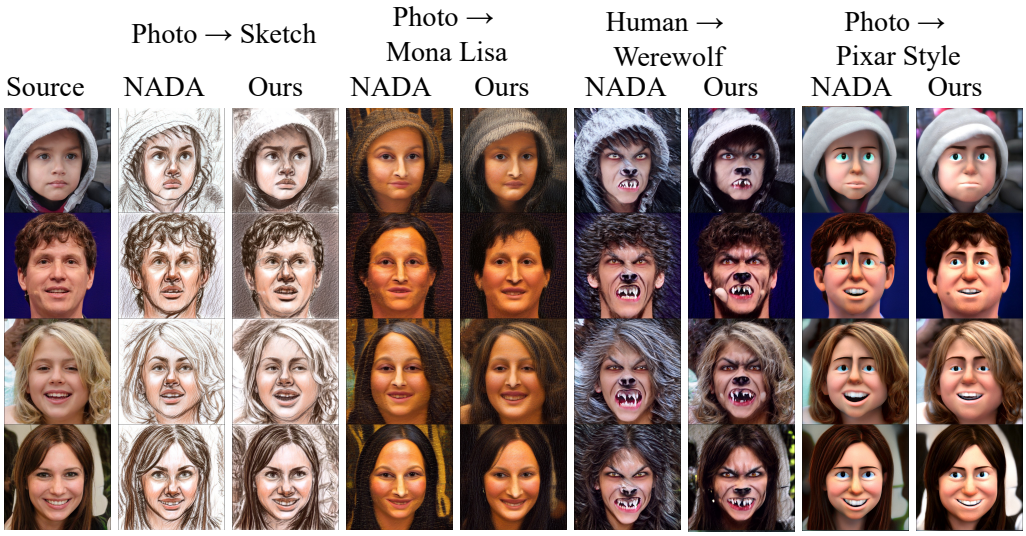

Figure 4: Visual results of our method and state-of-the-art StyleGAN-NADA Gal et al. (2021) for different domain adaptations. The top row shows the source text and target text description. The left column presents samples generated by source generator. Our method not only generates samples described by target text with high fidelity, but also produces diverse and identity consistent images corresponding to source domain.

loss(Zhu et al., 2021), the Mode Seeking(MS) loss(Mao et al., 2019), Perceptual Path Length(PPL) regularization(Karras et al., 2020) and Cross Domain Correspondence(CDC) regularization(Ojha et al., 2021). Detailed explanations of these regularization methods are in Appendix A.6. As shown in Figure 5, SCC, Within and MS regularization impose too strong regularization to target generator and restrict the domain specific attributes for target domain. Suffering from mode collapse problem, the generation results of PPL and CDC share the same pattern across different samples. In comparison, our method has a better balance between the diversity and stylization of target generator.

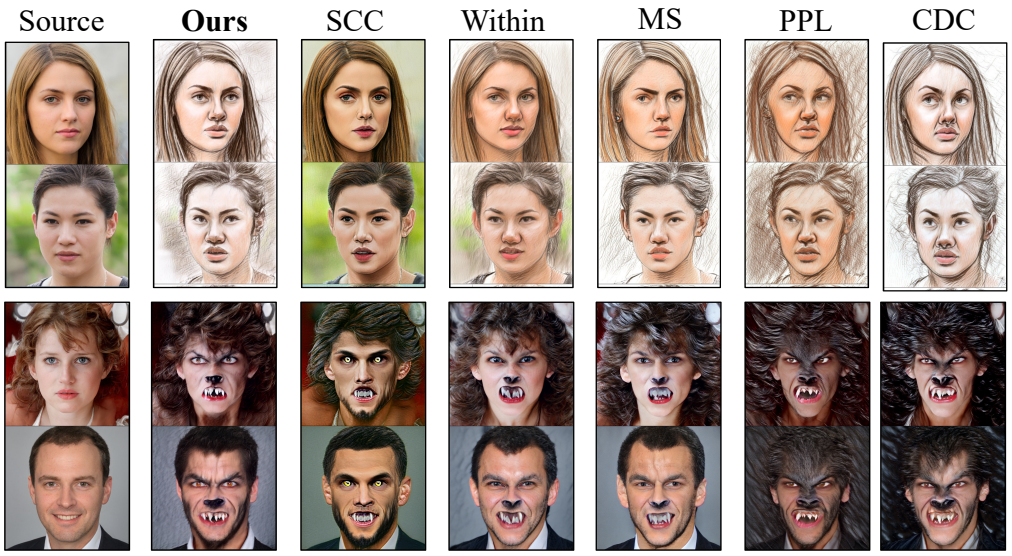

Figure 5: Comparison results of our spectral consistency regularization with other regularization methods. Detailed explanations of these methods are in Appendix A.6.

Quantitative Comparison Besides our proposed trace norm of Hessian Matrix in Equation 3, we also leverage the Perceptual Path Length(PPL) Karras et al. (2019) for quantitative diversity com-

parison. It measures the perceptually-based pairwise image distance Zhang et al. (2018) for a linear interpolation path in the latent path. The average PPL in the latent space $\mathcal{Z}$ is

$$PPL_{\mathcal{Z}} = \mathbb{E}\left[\frac{1}{\epsilon^2} d\left(G\left(\text{slerp}\left(\mathbf{z}_1, \mathbf{z}_2; t\right)\right), G\left(\text{slerp}\left(\mathbf{z}_1, \mathbf{z}_2; t+\epsilon\right)\right)\right)\right],$$

where $d(.,.)$ evaluates the perceptual distance between two images, and slerp denotes spherical interpolation since the input latent is normalized. The perceptual path length estimates the diversity of the generator via finite differences, and our method estimates the diversity analytically.

The quantitative results are shown in Table 1. The mentioned strong regularization methods including SCC, Within and MS have large values of PPL and Hessian trace at the cost of restricting adaptation effects. Compared with CDC and PPL, our method better preserve the diversity of source domain reflected in PPL and Hessian Trace. For all different target domains, our method outperforms previous state-of-the-art model StyleGAN-NADA for both PPL and Hessian trace. This benefits from that our method can alleviate the mode collapse problem and generate more diverse images without restricting the style adaptation results.

Table 1: Comparison results for diversity by PPL and Hessian trace between different regularization methods. The PPL and Hessian trace for the pretrained source generator is 419.22 and 0.309.

Results	Photo → Sketch		Photo → Mona Lisa		Human → Werewolf		Photo → Pixar	
	PPL	Trace	PPL	Trace	PPL	Trace	PPL	Trace
SCC	547.34	0.526	485.70	0.521	428.45	0.343	440.17	0.535
Within	378.25	0.090	363.17	0.191	311.34	0.116	345.31	0.167
MS	466.99	0.167	331.06	0.233	352.61	0.191	377.88	0.358
PPL	241.50	0.028	209.09	0.061	297.74	0.062	300.19	0.063
CDC	348.01	0.062	259.92	0.079	351.73	0.089	299.49	0.111
NADA	323.25	0.061	281.59	0.098	302.91	0.101	343.54	0.112
Ours	463.57	0.116	321.43	0.140	383.19	0.137	353.80	0.181

4.2 Ablation Study

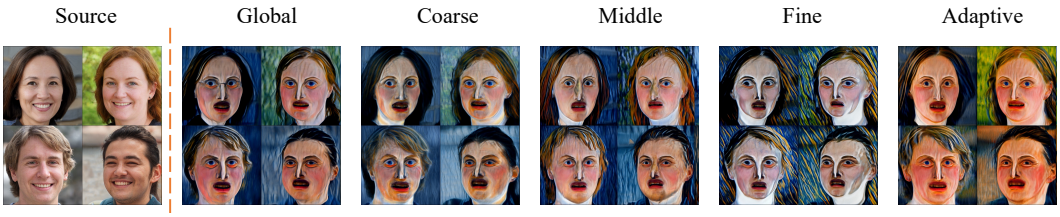

Figure 6: Results of different regularization strategies for the domain adaptation from photo to Edvard Munch paintings.

Granularity Adaptive Regularization. In Figure 6, we demonstrate generated results with different regularization strategies applied in $\mathcal{W}+$ space in Section 3.4. Regularization to the coarse scale will preserve the structure of source image but the diversity of fine features like hair will lose. The global regularization performs similarly to coarse regularization since the coarse features dominate the diversity of generator. Only applying regularization to the fine scale will generate high-frequency textures and the structure characteristic like necks will collapse. In comparison, our proposed granularity adaptive regularization both preserves the diversity of source domain in all scales and matches the styles of target domain.

Strength of regularization. In Figure 7, we show the generated samples with a linear interpolated loss weight $\lambda_{spectral}$. We can observe that with increasing $\lambda_{spectral}$, the generated samples maintain more diversity of source domain, and they also illustrate the most significant characteristic of target domain.

Choices of metric space. We conduct experiments about the metric space of spectral consistency regularization with different feature encoder $\phi(x)$ that evaluates the distance between image samples. Besides CLIP Radford et al. (2021) image encoder in our method, we also leverage

Figure 7: Generation results of linearly interpolated $\lambda_{spectral}$ in regularization.

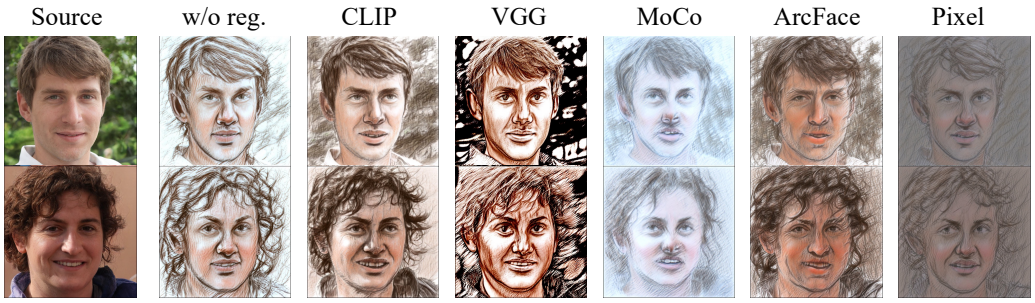

Figure 8: The generated results with spectral consistency regularization under the metric spaces defined by different encoders.

VGG Simonyan & Zisserman (2015) which was commonly used in style transfer Gatys et al. (2016), MoCo He et al. (2020) of contrastive representation learning, ArcFace Deng et al. (2019) for face recognition and the plain pixel space. As shown in Figure 8, compared to other feature encoders, the spectral consistency regularization with CLIP encoder shows best performance for preserving the identity of source image.

4.3 Applications

Image-to-Image Translation We combine the adapted generator in target domain with a GAN inversion encoder to implement image-to-image translation. Given a real-world image, we invert it to the latent code in $\mathcal{W}$ space via an e4e encoder Tov et al. (2021), which is then fed to target generator to produce target image. As shown in Figure 12, our method can achieve high-quality image translation and preserve the identity of source image for different domains.

Image Editing In Figure 13, we demonstrate the image editing results performed on the target domain. We leverage the meaningful directions found by InterfaceGAN Shen et al. (2022) to edit target images. We can observe that the editing directions from source domain still apply to target domains, which proves that the target domain preserves the semantic distribution of source domain.

5 Conclusion

In this paper, we propose the spectral consistency regularization for text-driven domain adaptation. The key insight of our method is to build a quantitative diversity estimator to preserve the intra-domain diversity of source generator without restricting the adaptation of target style. We also introduce an adaptive regularization strategy for granularity-flexible adaptation. The experiments demonstrate our method greatly improves the generation results for a wide range of target domains.

REFERENCES

Katherine Crowson, Stella Rose Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *ArXiv*, abs/2204.08583, 2022.

Jiankang Deng, J. Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, 2021.

Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *ArXiv*, abs/2106.14843, 2021.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ArXiv*, abs/2108.00946, 2021.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *ArXiv*, abs/2004.02546, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 18:1059–1076, 1989.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *ArXiv*, abs/2111.03186, 2021.

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1429–1437, 2019.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *ArXiv*, abs/2002.10964, 2020.

Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2750–2758, 2019.

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. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10738–10747, 2021.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2065–2074, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9240–9249, 2020.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2004–2018, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40:1 – 14, 2021.

Binxu Wang and Carlos R. Ponce. The geometry of deep generative image models and its applications. *ArXiv*, abs/2101.06006, 2021.

Jiayu Xiao, Liang Li, Chaofei Wang, Zhengjun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. *ArXiv*, abs/2203.04121, 2022.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv*, abs/2206.10789, 2022.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5908–5916, 2017.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yabo Zhang, Mingshuai Yao, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Towards diverse and faithful one-shot adaption of generative adversarial networks. *ArXiv*, abs/2207.08736, 2022.

Peihao Zhu, Rameen Abdal, John C. Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *ArXiv*, abs/2110.08398, 2021.

A APPENDIX

A.1 TRAINING DETAILS

We use the StyleGANv2 Karras et al. (2020) generator pretrained on FFHQ as the source generator. During domain adaptation, we optimize all the parameters of generator except for the mapping network and toRGB layers. We use the Adam Kingma & Ba (2015) optimizer with learning rate 0.001. The $\lambda_{spectral}$ is set to 1.0. For most target domains, only 300 iterations are required to achieve convergence. Following CLIP, we concatenate 79 manually designed prompts like "a photo of a..." with provided target domain description and feed them to text encoder of CLIP to get the embeddings of target domain.

A.2 CHOICE OF INPUT SPACE

Denote the mapping network $\mathcal{Z} \rightarrow \mathcal{W}$ as $f(z;\theta)$, the synthesis network $\mathcal{W} \rightarrow \mathcal{I}$ as $g(w;\phi)$, the generator network can be seen as $g(f(z;\theta);\phi)$. Derived from Equation 3 and the chain rule, the Hessian matrix of generator can be calculated as $H_{g\cdot f} = (J_g J_f)^T (J_g J_f) = J_f^T (J_g^T J_g) J_f = J_f^T H_g J_f$. There are two choices for spectral consistency regularization, which are $H_{g\cdot f}$ and H_g. We find regularizing H_g is unstable and easily cause artifacts since the density of W space is not ensured compared to Z space as a standard normal distribution.
As shown in Equation 6 in paper, we use Jacobian-Vector product operation to estimate the Hessian Trace. Since the mapping network is frozen during training, we precompute $e = J_f v$ (e.g. not create compute graph for backpropogation) and only optimize $e^T J_g^T J_g e$ during backpropogation. In the image sampling process, there involves the style-mixing technique, which feeds concatenated w codes generated from different z codes to the synthesis network. To integrate this technique, we use W+ space instead of W space for spectral consistency regularization. Furthermore, the W+ space provide the ability for granularity adaptive regularization as shown in Section 3.4 of paper.
In summary, we regularize the Hessian Matrix $g(f(z;\theta);\phi)$ with respect to z. Compared to Z+ space, we use W+ space for Jacobian-Vector product optimization to avoid overhead compute cost brought by mapping network during training.

A.3 DISCUSSION WITH PREVIOUS WORKS

Related to our work, Wang & Ponce (2021) also exploit the generative models from the geometric point of view. We both leverage the Hessian Matrix to analyze the characteristic of generative models. Here we explain the differences between these two works in details. First, the aim of Wang & Ponce (2021) is to find interpretable directions in GAN latent space, which can be grouped into the task of unsupervised GAN editing which also includes Härkönen et al. (2020) and Shen & Zhou (2021). Instead, we try to solve the mode collapse problem in generative domain adaptation with spectral consistency regularization. To the best of our knowledge, we are the first to deploy Hessian Matrix to the domain adaptation problem. Second, Wang & Ponce (2021) leverage the top eigenvectors in Hessian Matrix to find the interpretable directions, while what we utilize is the trace norm of Hessian Matrix, e.g. the sum of all eigenvalues. Third, Wang & Ponce (2021) use Lanczos iteration for numerical calculation of top eigenvectors, and our work use Hutchinson's method to estimate the trace norm.

A.4 ADDITIONAL QUALITATIVE RESULTS

In Figure 9 and Figure 10, We demonstrate our results for text-driven generative domain adaptation from the pretrained generator on dogs and cars datasets.

A.5 ADDITIONAL COMPARISON RESULTS WITH OTHER REGULARIZATION METHODS

We demonstrate more comparison results with other regularization methods in Figure 11. It's obvious in Figure 11 that while MS and Within are able to preserve the attributes of source image, they

Figure 9: Text-driven domain adaptation for pre-trained generator of dogs.

Figure 10: Text-driven domain adaptation for pre-trained generators of cars.

often restrict style adaptation too much, especially in the last row that the faces generated by MS and Within are more like photos instead of paintings. Different from that, our methods achieve a suitable balance between the effects of adaptation and attributes preservation.

A.6 Details of other regularization methods

Selective Cross-modal Consistency (SCC) was proposed to select and retain the domain-sharing attributes in $\mathcal{W}+$ space(Zhang et al., 2022). First, $G_s(z)$ and $G_t(z)$ are inverted into latent codes w_s and w_t in $\mathcal{W}+$ space with a pre-trained inversion model for each iteration. Then, we calculate the differences Δw between the centers of a queue of $\mathcal{W}+$ latent codes $\mathcal{X}_s$ and a queue of $\mathcal{W}+$ latent codes $\mathcal{X}_t$, where $\mathcal{X}_s$ and $\mathcal{X}_t$ are dynamically updated with w_s and w_t. The SCC loss is computed as:

$$\mathcal{L}_{scc} = \left\| \text{mask}(\Delta \boldsymbol{w}, \alpha) \cdot (\boldsymbol{w}_B - \boldsymbol{w}_A) \right\|_1,$$

where α represents the proportion of preserved attributes $\text{mask}(\Delta \boldsymbol{w}, \alpha)$ determines which channels to be retained. Let $|\Delta \boldsymbol{w}_{s_{\alpha N}}|$ be the αN-th largest element of $|\Delta w|$, and each dimension of $\text{mask}(\Delta \boldsymbol{w}, \alpha)$ is calculated as:

$$\text{mask}(\Delta \boldsymbol{w}, \alpha)_i = \begin{cases} 1 & |\Delta \boldsymbol{w}_i| < |\Delta \boldsymbol{w}_{s_{\alpha N}}| \\ 0 & |\Delta \boldsymbol{w}_i| \geq |\Delta \boldsymbol{w}_{s_{\alpha N}}| \end{cases}$$

Mode Seeking(MS) Loss was proposed to alleviate mode collapse in image generation(Mao et al., 2019). The mode seeking regularization term directly maximizes the ratio of the distance between $G(z_1)$ and $G(z_2)$ with respect to the distance between z_1 and z_2,

$$\mathcal{L}_{\text{ms}} = \frac{d_{\mathbf{I}}\left(G\left(\mathbf{z}_1\right), G\left(\mathbf{z}_2\right)\right)}{d_{\mathbf{z}}\left(\mathbf{z}_1, \mathbf{z}_2\right)}$$

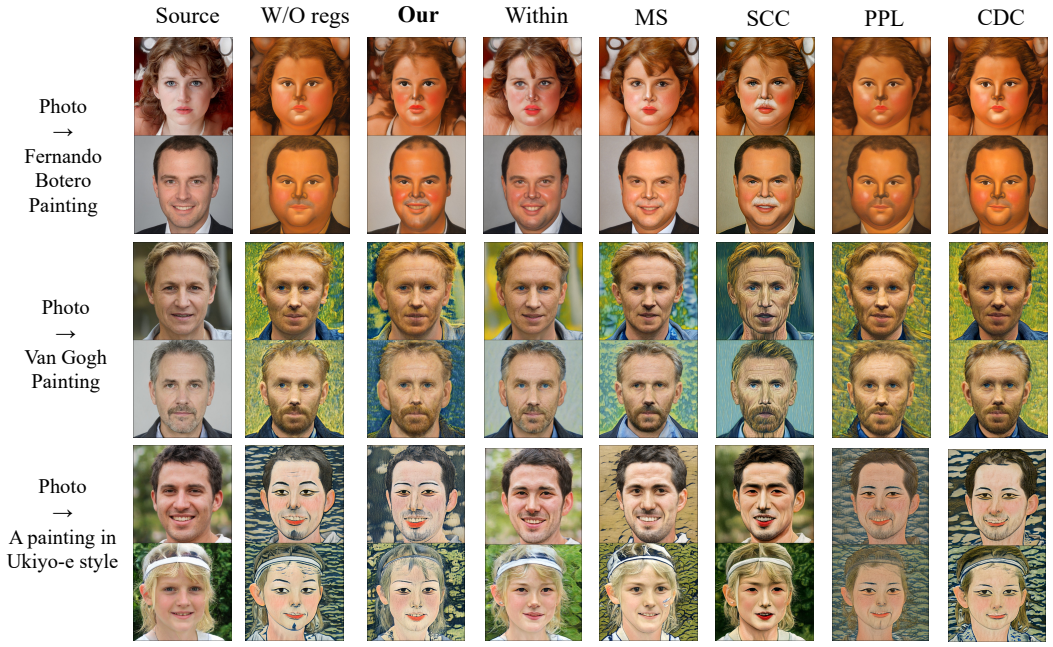

Figure 11: Comparison results of our spectral consistency regularization with other regularization methods. The second column are generated results without any regularization method.

where d_I and d_z denotes the distance metric. Here we use the CLIP cosine distance for d_I and Euclidean distance for d_z.

Perceptual Path Length(PPL) regularization was proposed to encourage that a fixed-size step in $\mathcal{W}$ results in a non-zero, fixed-magnitude change in the image(Karras et al., 2020). The PPL regularizer is formulated as:

$$\mathbb{E}_{\mathbf{w},\mathbf{v}\sim\mathcal{N}(0,\mathbf{I})}\left(\left\|\mathbf{J}_{\mathbf{w}}^{T}\mathbf{v}\right\|_{2}-a\right)^{2}$$

where $\mathbf{J}_{\mathbf{w}} = \partial G(\mathbf{w})/\partial \mathbf{w}$. The constant a is set dynamically during optimization as the running exponential moving average of the lengths $\left\|\mathbf{J}_{\mathbf{w}}^{T}\mathbf{v}\right\|$.

Within Domain Consistency Loss was proposed to preserve the semantic information that is not related to the domain gap between source domain and target domain(Zhu et al., 2021). Let $v_s = E_I(G_s(w_1)) - E_I(G_s(w_2))$ be a vector between two samples in source domain under CLIP space. Let $v_t = E_I(G_t(w_1)) - E_I(G_t(w_2))$ denote the corresponding vector in target domain. The within domain consistency loss is formulated as:

$$L_{clip\_within} = 1 - sim(v_s, v_t)$$

In one-shot domain adaptation, the $G_t(w_2)$ and $G_s(w_2)$ are replaced with provided reference image of target domain and its corresponding inversion image in source domain.

Cross Domain Correspondence(CDC) regularization was proposed to preserve the relative distance in the source domain during adaptation(Ojha et al., 2021). First, sample a batch of $N+1$ noise vectors $\{z_n\}_0^N$, and use their pairwise similarities in feature space to construct N-way probability distributions for each image. The probability distribution for the i^{th} noise vector, for the source and adapted generators is given by,

$$y_i^{s,l} = \text{Softmax}\left(\left\{\text{sim}\left(G_s^l\left(z_i\right), G_s^l\left(z_j\right)\right)\right\}_{\forall i\neq j}\right)$$
$$y_i^{s\to t,l} = \text{Softmax}\left(\left\{\text{sim}\left(G_{s\to t}^l\left(z_i\right), G_{s\to t}^l\left(z_j\right)\right)\right\}_{\forall i\neq j}\right)$$

where sim denotes the cosine similarity between generator activations at the l^{th} layer. The adapted model is encouraged to have similar distributions to the source, across layers and images instances

by using KL-divergence:

$$\mathcal{L}_{\text{dist}}\left(G_{s\to t}, G_s\right) = \mathbb{E}_{\{z_i\sim p_z(z)\}} \sum_{l,i} D_{KL}\left(y_i^{s\to t,l} \| y_i^{s,l}\right)$$

Figure 12: Image translation of real-world images to different target domains. Each column shows a target domain, and the top row is the text description for target domain. The transferred images represent both target style and the identity of source image.

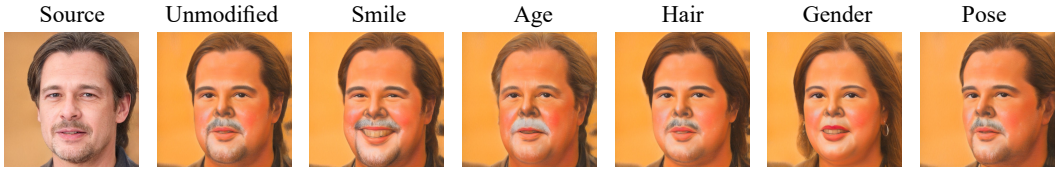

Figure 13: Editing images in target domain for real-world images. The top row shows the edited attributes.

A.7 Applications

As mentioned in Section 4.3, we conduct experiments for applications including image-to-image translation and image editing. In Figure 12, we show the image translation results from real-world photo to different target domains. We can observe that our translated images both preserve that identity of source image and satisfy the style of target domain. In Figure 13, we illustrate the image editing results of target domain via meaningful directions for the source domains. The editing results are consistent with the specified editing attributes, which proves that our adaptation model can preserve the semantic distribution of source domain.

A.8 The relation between Hessian Trace and Diversity

Considering the mapping from standard normal distribution $z \sim \mathcal{N}(0, I)$ to general multivariate normal distribution $y \sim \mathcal{N}(\mu, \Sigma)$ with the generator as a linear function $G(z) = Az + b$, which has $\mu = b, \Sigma = AA^T$. The linear assumption of generator holds when we consider a small neighborhood around z, i.e. $\lim \Delta z \to 0$.

The variance of y can be computed as:

$$\begin{aligned} Var(y) &= E[\|y - E[y]\|_2^2] \\ &= E[\sum_{i=1}^{n}(y_i - E[y_i])^2] \\ &= \sum_{i=1}^{n} E[(y_i - E[y_i])^2] \\ &= \sum_{i=1}^{n} Var(y_i) \\ &= \sum_{i=1}^{n} \Sigma_{ii} \\ &= Trace(\Sigma) = Trace(AA^T) \end{aligned}$$

On the other side, the $d^2(G(z), G(z + \Delta z))$ can be expanded as follows:

$$\begin{aligned} d^2(G(z), G(z + \Delta z)) &= \|G(z) - G(z + \Delta z)\|_2^2 \\ &= \|Az + b - (A(z + \Delta z) + b)\|_2^2 \\ &= \|A\Delta z\| \\ &= \Delta z^T A^T A \Delta z \end{aligned}$$

such the Hessian Matrix H of Δz with respect to $d^2(G(z), G(z + \Delta z))$ equals to $A^T A$, e.g. $H = A^T A$. Combining above two equations with $Trace(AA^T) = Trace(A^T A)$, we have $Var(y) = Trace(H)$, which means that the Hessian Trace for every sample in target distribution reflects the variance and diversity of this distribution. If the Hessian Trace is small, the target distribution only spans a small region in space. This is consistent to the mode collapse problem in generative models.

A.9 Interpolation Results

We provide examples for latent interpolation and cross-model interpolation results in Figure 14 and Figure 15. In Figure 14, the latent interpolation results present a smooth transition between two different generated samples, and interpolated samples still keep high fidelity. We also linearly interpolate the parameters of models from two target domains to get cross-model interpolation results. In Figure 15, the generated images demonstrate the style transition between different domains with identity preservation.

Figure 14: Latent interpolation results for different target domains.

Figure 15: Cross-model interpolation results for Pixar, Werewolf and Sketch.

A.10 Visualization of Hessian Trace During Adaptation

In Figure 16 and Figure 17, we show the Hessian Trace of generator during the adaptation process compared with StyleGAN-NADA. Different from StyleGAN-NADA which decreases rapidly, our method can preserve the diversity of generator at the late stage of training.

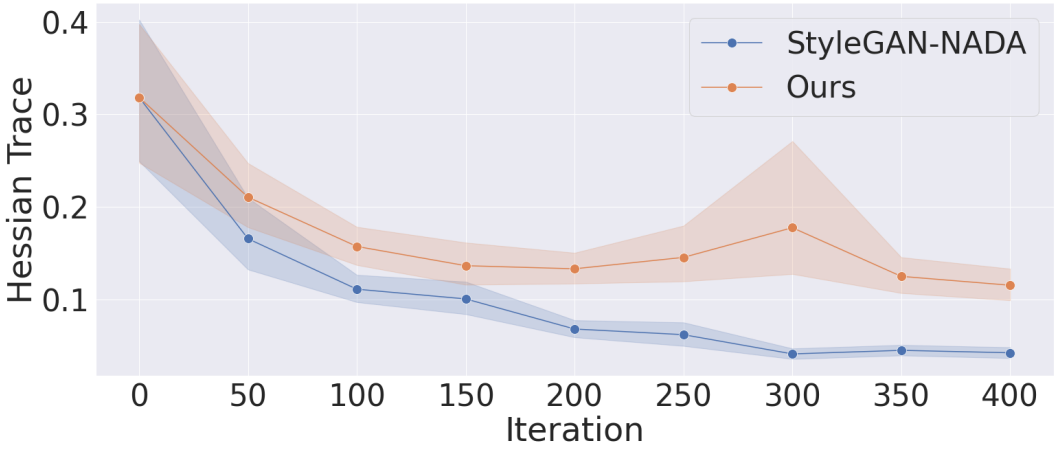

Figure 16: Illustration for the Hessian Trace of generator during the adaptation process from photo to sketch.

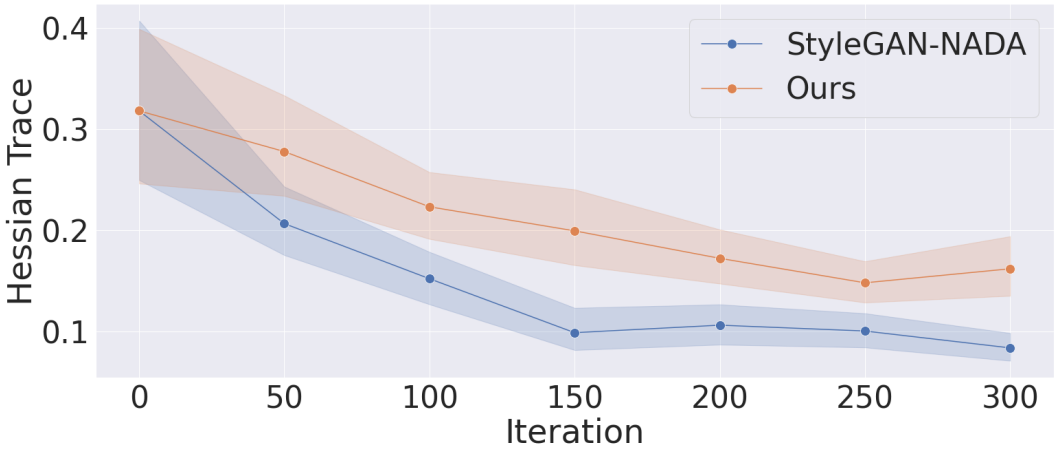

Figure 17: Illustration for the Hessian Trace of generator during the adaptation process from human to werewolf.