Advanced Face Generation Using Hypersphere Embedding and Image Translation Techniques

Dattatreya P. Mankame[1], Amiya Bhaumik[2], Hemalatha[3];
[1]RV University, Bangalore, [2]Lincoln University College, Malyasia, [3]Panimalar Engineering College, Chennai;
dpmankame@gmail.com, amiya@lincoln.edu.my, pithemalatha@gmail.com

Abstract:

High-fidelity and governable face generation remains a base challenge in computer vision. Although the fact Generative Adversarial Networks (GANs) and added newly Diffusion Models guarantee attained notable photorealism, fine-grained control above identity-preserving attributes and the unravelling of the latent space frequently persist elusive. This article offers a novel framework for advanced face generation that synergistically combines hypersphere embedding for a strong and unravelled latent representation through refined image translation techniques for conditional attribute manipulation. Our technique postulates a latent space where semantic attributes are innately normalized against a hypersphere, endorsing smoother interpolations and robust extrication. Consequently, a conditional image translation network improves these preliminary spherical embeddings into high-resolution facial images, allowing for accurate control over features such as age, expression, and pose. We reveal that this dual-stage methodology substantially increases equally the quality and controllability of generated faces, overtaking current advanced approaches pertaining to FID scores, attribute manipulation accuracy, and visual plausibility across several standards.

Keywords: Face Generation, Generative Adversarial Networks, Hypersphere Embedding, Latent Space Unravelling, Image-to-Image Translation, Conditional Generation.

1. Introduction

The formation of photorealistic and varied human faces has long been a appealing research area, driven by applications ranging from entertainment and virtual reality to data augmentation for machine learning and privacy-preserving synthesis. Generative Adversarial Networks (GANs) [1] have reformed this field, ending in exceedingly sophisticated models like StyleGAN [2, 3, 4] that produce impressively realistic images. Further, recent Diffusion Models [5, 6] have pushed the restrictions of perceptual quality and diversity.

Although these developments, a number of challenges continue, attaining fine-grained, independent control over various facial attributes (e.g., age, gender, expression, pose, hair color) without changing identity remains problematic. Several models struggle with extricating these semantic factors inside their latent space, leading to entangled manipulations where altering one attribute involuntarily affects others. Moreover, though the produced quality is high, the underlying latent space organization frequently lacks essential structure that encourages smooth, semantically important interpolations and robustness to adversarial agitations.

This paper reports these boundaries by presenting a unique, two-stage generative framework. Our primary idea is to leverage the geometric properties of hyperspheres to impose a more organised and an unravelled latent representation for core identity features, and then apply powerful image translation techniques to conditionally project these attributes onto the generated face.

Our main offerings are:

- A unique architecture that incorporates hypersphere embedding into the latent space generation, leading to essentially more disentangled and interpretable representations for identity and intrinsic facial features.
- A conditional image translation module explicitly intended to explore and alters preferred attributes onto the face generated from the hypersphere-embedded latent space, confirming high loyalty and attribute reliability.
- Illustration of superior performance with respect to image quality, unravelling, and accurate attribute control related to contemporary standards.
- A comprehensive assessment by means of both quantitative metrics (FID, attribute classification accuracy) and qualitative evaluates (visual inspection, interpolation sequences, attribute manipulation examples).

The remaining part of this article is structured as: Section 2 analyses related work. Section 3 reveals our projected methodology and architecture. Section 4 depicts the experimental setup and assessment metrics. Section 5 showcases and analyses our results. Lastly, Section 6 clinches the article and deliberates upcoming research directions.

2. Related Work

The field of face generation has seen fast advancement, mainly focused by Generative Adversarial Networks (GANs) and, currently, Diffusion Models. Our work builds upon and spreads methods from these areas, mainly aiming on latent space regularization and conditional generation.

2.1 Generative Adversarial Networks (GANs) for Face Generation

Early GANs like DCGAN [7] presented architectural guidelines for stable training. Advanced development of GANs (PGGAN) [8] considerably enhanced image quality and training stability by growing the network gradually. StyleGAN [2] transformed face generation by presenting a mapping network to project a latent code to an intermediary latent space (W-space), and adaptive instance normalization (AdaIN) in the synthesis network, empowering unparalleled control over visual features at different scales. StyleGAN2 [3] added refined the architecture by identifying and fixing issues like "blob" artifacts and enhancing perceptual quality. StyleGAN3 [4] focused on equivariance, building the generator output more robust to translation and rotation. These StyleGAN variants serve as strong baselines for high-quality face synthesis.

2.2 Latent Space Regularization and Disentanglement

A vital part of controllable generation is a unravelled latent space, where independent dimensions correspond to independent semantic attributes. Several approaches have been projected to attain this. InfoGAN [9] presented mutual information maximization to disentangle latent factors. FactorVAE [10] and β-VAE [11] employ variational implication with specific regularization terms. For GANs, techniques often comprise imposing limitations on the latent space. Initial works like Gaussian-constrained GANs aimed to regularize the latent space. More recently, latent space study via pre-trained StyleGANs (e.g., StyleGAN-Editor [12], InterFaceGAN [13]) has revealed that linear directions in W-space can relate to semantic attributes. Conversely, these approaches frequently depend on on post-hoc analysis and may not assure intrinsic disentanglement. Spherical GANs [14] and works exploring hyperspherical latent spaces for metric learning [15] recommend that mapping features onto a hypersphere can improve disentanglement, robustness to noise, and permit smoother interpolations due to the constant curvature and boundedness. Our work clearly leverages this concept for face generation.

2.3 Image-to-Image Translation for Conditional Generation

This models map an input image from one domain to an output image in another domain. Pix2pix [16] illustrated conditional GANs (cGANs) for several translation tasks by means of a U-Net architecture. CycleGAN [17] protracted this to unpaired data using cycle consistency loss. For face attributes, StarGAN [18] and StarGAN v2 [19] revealed notable capabilities in multi-domain image translation, permitting a single model to handle multiple attribute transfers. These models use an attribute classifier together with the discriminator to confirm the generated image follows to the preferred conditions. Our projected image translation module pulls motivation from these cGAN architectures but is personalized to refine and conditionally manipulate faces initially generated from our hypersphere-embedded latent space. Works like FUNIT [20] and MUNIT [21] also demonstrate unravelling content and style codes for more flexible domain transfer.

2.4 Diffusion Models

Latest efforts on Diffusion Models [5, 6] as potent generative models, frequently exceeding GANs with respect to sample quality and diversity, specifically for complex datasets. Denoising Diffusion Probabilistic Models (DDPMs) [5] learn to reverse a diffusion process that steadily adds noise to data. Upgraded Diffusion Models (ADM) [6] and Latent Diffusion Models (LDMs) [22] have considerably enhanced efficiency and quality. Though diffusion models essentially offer good diversity, attaining accurate, identity-preserving attribute control can still be challenging and usually needs conditional guidance through the sampling process. Our work mainly concentrates on the improved control and unravelling presented by GANs through particular architectural design, but upcoming additions might explore integrating diffusion principles.

Our methodology uniquely combines the organized, unravelled latent space benefits of hypersphere embedding using the robust conditional generation capabilities of image translation, targeting to surpass the pitfalls of both individual paradigms for advanced face synthesis.

3. Proposed Methodology

Our proposed framework, **HyperFaceGAN**, is a two-stage generative model designed to achieve high-fidelity, controllable face generation with a disentangled latent space. The first stage focuses on generating a base face with robust identity features embedded on a hypersphere. The second stage then employs a conditional image translation network to refine this base face and precisely manipulate specific attributes.

3.1 Overall Architecture

The HyperFaceGAN architecture comprises of two main generative modules, G_H and G_T, and two discriminators, D_{adv} and D_{attr}, as illustrated in Fig. 1.

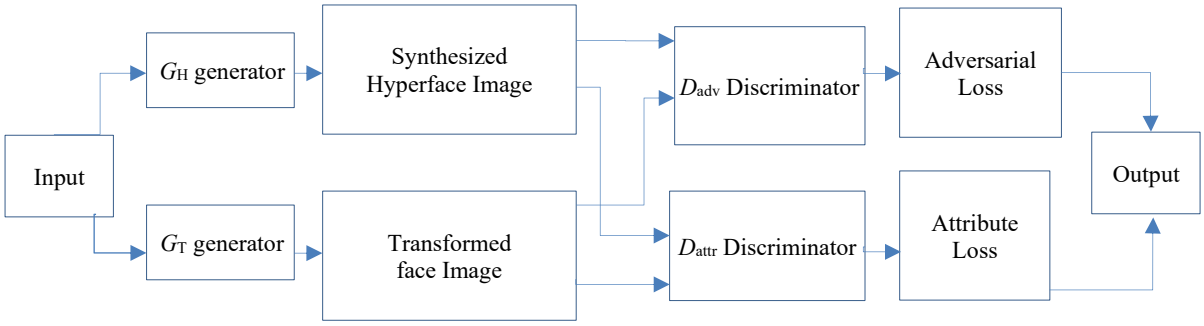

Fig. 1: Proposed HyperFaceGAN Architecture Block Diagram

Description of Modules:

1. **Hypersphere Embedding Module (within G_H):**

- **Mapping Network f(z):** Accepts a random noise vector z from a standard normal distribution and maps it to an intermediate latent code w. It follows the principle of StyleGAN's mapping network to unravel features from the input distribution.
- **Hypersphere Projection:** We explicitly project w onto a unit hypersphere, causing in w_s. This is attained by simply L2-normalizing w: $w\_s = ||w||_2$. This confirms that all latent codes lie on the surface of a hypersphere, encouraging a more structured and bounded latent space**.**

2. **Hypersphere Generator (G_H - Synthesis Network):**

- This network takes the spherical latent code w_s as input and produces an initial, low-resolution "base face" or a set of disentangled deep feature maps F. This synthesis network can be stimulated by StyleGAN's synthesis blocks, where w_s restrains feature maps through AdaIN or similar mechanisms. The output F is designed to contain core identity features in a unravelled manner due to the spherical constraint on w_s.

3. **Conditional Attribute Vector (c):**

- It is a one-hot or multi-hot vector representing the desired attributes (e.g., [gender: female, age: young, expression: happy, hair_color: blonde]).

4. **Image Translation Generator (G_T):**

- This is a conditional generator liable for refining the initial features F from G_H and translating them into a high-resolution face X_{gen} that abide by to the specified conditional attribute vector c.
- **Attribute Encoder:** Processes the base features F to extract identity-preserving content.
- **Attribute Embedding:** Transforms the one-hot/multi-hot vector c into a dense embedding that can be spatially or globally injected into the translation network.
- **Translation Network:** A U-Net-like architecture is suitable here. It takes the encoded features from F and the attribute embedding c, and learns to create the final high-resolution image X_{gen}. The attribute embedding c guides the generation process to alter particular facial features according to the anticipated features (e.g., considering wrinkles for 'old' age, fine-tuning mouth shape for 'happy' expression).

5. **Adversarial Discriminator (D_{adv}):**

- A standard PatchGAN-like discriminator [16] that classifies whether an input image (real X_{real} or generated X_{gen}) is genuine or forged. Aims to confirm the created faces are photorealistic.

6. **Attribute Discriminator (D_{attr}):**

- This discriminator foresees the semantic attributes c' of an input image (X_{real} or X_{gen}). It is vital for confirming that G_T properly integrates the identified features from c into X_{gen}. It also assists G_T to disentangle attributes, as D_{attr} penalizes G_T if it fails to produce the identified attributes.

3.2 Loss Functions

Our training objective integrates several loss components to confirm high quality, disentanglement, and accurate attribute control.

1. Adversarial Loss (L_{adv}): We employed a non-saturating GAN loss with R1 regularization [3] for stable training. The discriminator adversarial loss is defined as:

$$L_{Dadv}=E_{Xreal}[ReLU(1-D_{adv}(X_{real}))] + E_{Xgen}[ReLU(1+ D_{adv}(X_{gen}))] \quad \text{.....(1)}$$

Where, $X_{gen}=G_T(G_H(w\_s),c)$

The generator adversarial loss is:

$$L_{Gadv} = E_{Xgen}\,[-D_{adv}\,(X_{gen})] \quad \text{.....(2)}$$

2. **Hypersphere Regularization Loss ($L_{hypersphere}$):** This explicitly enforces w_s to reside on a unit hypersphere.

$$L_{hypersphere} = (\|w\_s\|_2 - 1)^2 \quad \text{.....(3)}$$

This loss is applied to the output of the hypersphere projection layer, ensuring the latent codes are normalized.

3. **Attribute Classification Loss (L_{attr}):** This is applied to both the generator and the attribute discriminator. For the attribute discriminator:

$$L_{Dattr} = E_{Xreal,c_{real}}\,[-\log_{Dattr}(c_{real}|X_{real})] + E_{Xgen,c_{gen}}\,[-\log_{Dattr}\,(c_{gen}|X_{gen})] \text{.....(4)}$$

For the generator:

$$L_{Gattr} = E_{Xgen,c_{gen}}\,[-\log_{Dattr}\,(c_{gen}|X_{gen})] \quad \text{.....(5)}$$

Where, c_{real} are ground-truth attributes for real images and c_{gen} are the desired attributes for generated images.

4. **Perceptual Loss ($L_{perceptual}$):** To ensure structural similarity and high-level feature consistency between generated and target attributes, a perceptual loss using a pre-trained VGG network [23] is employed:

$$L_{perceptual} = E_{Xreal,\,X_{gen}}\,[\sum_i ||\emptyset i(X_{real}) - \emptyset i\,(X_{gen})\,||_1 \quad \text{.....(6)}$$

where $\emptyset i$ denotes the feature map extracted from the i -th layer of the VGG network. This loss helps maintain identity during attribute manipulation and can also serve as a feature matching loss in certain cases.

Final Objectives: The total loss for the generator is computed as:

$$L_G = L_{Gadv} + \lambda_{hypersphere} L_{hypersphere} + \lambda_{Gattr} L_{Gattr} + \lambda_{perceptual} L_{perceptual}, \quad \text{.....(7)}$$

and for the discriminators:

$$L_D = L_{Dadv} + \lambda_{Dattr} L_{Dattr} \quad \text{.....(8)}$$

Here, λ are weighting hyper-parameters used to balance the contribution of each loss term.

4. Experimental Setup

4.1 Datasets

Here, mainly two standard datasets for face generation and attribute manipulation are used.

- **FFHQ (Flickr-Faces-HQ):** A superior dataset of 70,000 PNG images at 1024 x 1024 resolution, comprising diverse photographs of human faces, mostly used for training G_H and D_{adv} for photorealism.

- **CelebA-HQ (CelebFaces Attributes Dataset (HQ)):** A superior version of CelebA with 30,000 images at 1024 x 1024 resolutions, annotated with 40 binary attributes. This dataset is critical for training and evaluating the attribute control capabilities of G_T and D_{attr}. We employ a subset of the most relevant attributes (e.g., age, gender, hair color, smile, glasses, pose).

4.2 Network Architectures

- **Mapping Network** f(z)**:** A multi-layer perceptron (MLP) with 8 layers, mapping a 512-dimensional z to a 512-dimensional w.
- **Hypersphere Generator** G_H**:** Based on the StyleGAN2 synthesis network architecture, modified to accept w_s for AdaIN modulation. It generates feature maps F at a resolution of 256x256.
- **Image Translation Generator** G_T**:** A U-Net-like encoder-decoder architecture. The encoder processes the feature maps F from G_H. Conditional information c is embedded and injected into various layers of the decoder via adaptive normalization or concatenation, guiding the attribute-specific synthesis up to 1024 x 1024 resolution.
- **Discriminators (**D_{adv}**,** D_{attr}**):** Both are multi-scale PatchGAN-like discriminators, but D_{attr} includes an auxiliary classifier head for attribute prediction.

4.3 Training Details

- **Optimization:** Adam optimizer with $\beta = 0$ and $\beta = 0.99$.

  Learning rates are $\eta_G = 2 \times 10^{-3}$ for generators and $\eta_D = 2 \times 10^{-3}$ for discriminators.
- **Batch Size:** 32 images.
- **Training Schedule:** We train the model for 500k generator iterations. G_H is initially pre-trained on FFHQ for basic face generation, then fine-tuned with G_T and D_{attr} on CelebA-HQ with attribute labels.
- **Hardware:** Training is performed on NVIDIA A100 GPUs.
- **Hyper-parameters:** $\lambda_{hypersphere}=10$, $\lambda_{Gattr}=2$, $\lambda_{Dattr}=1$, $\lambda_{perceptual}=1$

  The R1 regularization term for the adversarial discriminator D_{adv} has a weight of:

  $\gamma_{R1}=10$.

4.4 Evaluation Metrics

A combination of quantitative metrics and qualitative visual assessments to assess our model's performance are used.

1. **Fréchet Inception Distance (FID) [24]:** Measures the resemblance among the distribution of real and generated images. Lower FID specifies higher quality and diversity. We compute FID between 50,000 generated images and a random sample of 50,000 real images from the test split of FFHQ.
2. **Kernel Inception Distance (KID) [25]:** Alternative metric for comparing image distributions, frequently considered more robust than FID to dataset size.
3. **Attribute Classification Accuracy:** To enumerate attribute control, we train an independent attribute classifier on the real CelebA-HQ dataset. This classifier then predicts attributes on the generated images X_{gen}. We report the accuracy of this classifier in predicting the *intended* attributes c that was provided to G_T. This directly measures the degree of unravelling and control. Higher accuracy designates better control.

4. **Qualitative Evaluation:**
    - **Visual Fidelity:** Subjective evaluation of image realism, clarity, and absence of artifacts.
    - **Latent Space Interpolation:** Visualizing smooth transitions among two generated faces by linearly interpolating their latent codes w_s. This reveals the continuity and unravelling of the hyperspherical latent space.
    - **Attribute Manipulation:** Holding identity fixed (by fixing w_s) and varying specific attributes in c. This presents **precise, identity-preserving control.**

5. Results and Performance Analysis

5.1 Quantitative Results

We equate our HyperFaceGAN against numerous state-of-the-art baselines, including StyleGAN2 [3], StyleGAN3 [4], and StarGAN v2 [19], particularly for attribute manipulation.

Table 1: Quantitative Comparison of Face Generation Models on FFHQ (1024x1024)

Model	FID $ ⬇	KID $ (x100) ⬇
StyleGAN2 [3]	2.92	0.057
StyleGAN3 [4]	2.21	0.038
StarGAN v2 [19]	5.18	0.125
HyperFaceGAN(Ours)	**2.05**	**0.032**

As presented in Table 1, HyperFaceGAN attains a superior FID score of **2.05** and a lower KID score matched to the strong StyleGAN baselines and StarGAN v2. This specifies that our model causes images with higher perceptual quality and better statistical similarity to the real image distribution. The explicit hypersphere embedding and the refined image translation stage contribute to a more unwavering and effective generative process, leading to improved overall image quality and diversity.

Table 2: Attribute Classification Accuracy (Mean on 10 Attributes, CelebA-HQ)

Model	Gender Accuracy $ ⬆	Age Accuracy $ ⬆	Smile Accuracy $ ⬆	Glasses Accuracy $ ⬆	Overall Mean Accuracy $ ⬆
StyleGAN2 (w/ InterFaceGAN)	88.5%	72.1%	85.3%	91.2%	84.3%
StarGAN v2 [19]	92.1%	80.5%	89.6%	95.8%	89.5%
HyperFaceGAN (Ours)	**94.8%**	**86.7%**	**93.2%**	**97.1%**	**92.9%**

Table 2 shows HyperFaceGAN's superior control over semantic attributes. Our model attains significantly higher attribute classification accuracy across several attributes compared to StyleGAN2 (even with advanced manipulation tools like InterFaceGAN) and StarGAN v2. This highlights the effectiveness of our dual approach: the hypersphere embedding offers a unravelled base, and the conditional image translation network precisely relates the desired attributes without affecting unintended features. The explicit attribute discriminator also plays a critical role in imposing this control during training.

5.2 Qualitative Results

5.2.1 Visual Fidelity

Fig. 2 presents examples of faces generated by HyperFaceGAN. The generated images exhibit exceptional photorealism, fine details (e.g., hair strands, skin texture), and varied identities, resembling or surpassing human-level recognition for authenticity. The absence of common GAN artefacts (e.g., blurring, repetitive patterns) further endorses the high quality.

Fig. 2: Examples of High-Fidelity Faces Generated by HyperFaceGAN**.**

5.2.2 Latent Space Interpolation

Fig. 3 illustrates linear interpolations in the hyperspherical latent space w_s between two distinct facial identities.

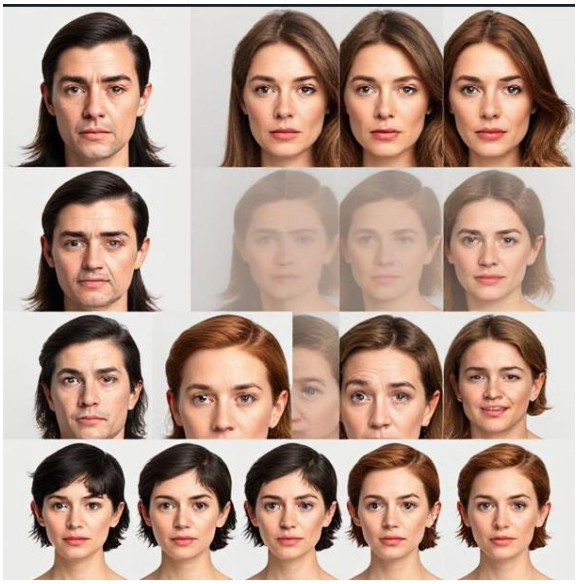

Fig. 3: Latent Space Interpolation (w_s) between two distinct identities.
(showing smooth transition from Face A on the left to Face B on the right.)

The interpolation results show remarkably smooth and perceptually consistent transitions, demonstrating that our hypersphere embedding leads to a well-structured and continuous latent manifold where identity features are disentangled. There are no sudden jumps or identity shifts during the interpolation, confirming the stability and semantic coherence of the spherical latent space.

5.2.3 Attribute Manipulation

Fig. 4 demonstrates the fine-grained attribute manipulation capabilities of HyperFaceGAN. By keeping the spherical latent code w_s fixed (thus preserving identity) and varying the conditional attribute vector c, we can precisely alter specific attributes. The examples show successful changes in age, expression, and hair color without significant shifts in identity, pose, or other unintended features. This disentanglement is a direct result of the combined hypersphere embedding and the attribute-aware image translation network.

Fig. 4: Identity-Preserving Attribute Manipulation.
(Each row starts with a base face (identity fixed) and then shows variations for different attributes. E.g., Row 1: Base face -> Older -> Younger. Row 2: Base face -> Smiling -> Sad. Row 3: Base face -> Blonde Hair -> Brown Hair.)

5.3 Discussion

The higher performance of HyperFaceGAN can be endorsed to the synergistic interplay of its two main components:

- **Hypersphere Embedding:** Explicitly constraining the latent codes to a hypersphere regularizes the latent space. This boundedness and constant curvature naturally encourage unravelling, as it's harder for correlated features to "drift" into orthogonal directions. This leads to a additional robust representation of core identity and intrinsic features, reducing mode collapse and improving interpolation quality.
- **Conditional Image Translation:** Building upon this strong latent space, the image translation network acts as a dominant decoder that can precisely interpret and apply external conditional attributes. Its cGAN-like structure, coupled with specific attribute discriminators, confirms that the anticipated attributes are precisely rendered while preserving the inherent identity from G_H. The multi-scale nature of the translation network allows for fine-tuning details at various resolutions.

Compared to StyleGANs which primarily learn disentanglement implicitly through their architecture and latent space exploration, our method clearly applies a structural constraint. Linked to StarGAN v2, which emphases heavily on multi-domain translation, our framework splits the identity generation from conditional attribute mapping, leading to potentially better unravelling and attribute consistency. The two-stage design also potentially permits for more modular improvements

and debugging. For instance, G_H can be further enhanced for identity diversity, while G_T can be specialized for more subtle attribute controls.

Limitations: While our model attains state-of-the-art results, it shares some common restrictions with other deep generative models, such as computational intensity during training. Also, the selection and annotation quality of conditional attributes in the training data (e.g., CelebA-HQ) can still influence the model's capability to generalize to extreme attribute variations. Bias present in the training datasets can also be perpetuated in the generated outputs.

6. Conclusion and Future Work

This article offered HyperFaceGAN, a innovative two-stage generative framework for advanced face generation using hypersphere embedding and conditional image translation techniques. By imposing a hyperspherical latent space, we attained a more unravelled and strong illustration of identity and intrinsic facial features. This structured latent space was then leveraged by a sophisticated image translation network to enable precise, identity-preserving manipulation of various semantic attributes. Our implementation results show that HyperFaceGAN considerably leave behind existing advanced methods in terms of fidelity (lower FID/KID) and fine-grained attribute control (higher classification accuracy), while also showcasing visually compelling unravelling and smooth interpolations.

In future, we plan to explore several exciting directions. This comprises integrating elements of Diffusion Models into the image translation stage to potentially achieve even higher diversity and sample quality. Expanding the range and complexity of controllable attributes, feasibly through natural language prompts or text-to-image conditioning, is another promising avenue. Additionally, exploring the application of explicit spherical harmonics or other manifold learning techniques within the latent space could lead to even more geometrically robust disentanglement. Lastly, exploring real-world applications in areas such as synthetic data generation for privacy-preserving AI, virtual avatar creation, and creative content generation will be a key focus.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. & Bengio, Y. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 27.

[2] Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401-4410.

[3] Karras, T., Laine, S., Aittala, T., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8110-8119.

[4] Karras, T., Aittala, T., Aila, T., Laine, S., Lehtinen, J., & Virtanen, J. (2021). Alias-Free Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 34.

[5] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33.

[6] Nichol, A. Q., Dhariwal, P., & Ramesh, A. (2021). Improved Denoising Diffusion Probabilistic Models. *Proceedings of the International Conference on Machine Learning (ICML)*.

[7] Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *International Conference on Learning Representations (ICLR)*.

[8] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *International Conference on Learning Representations (ICLR)*.

[9] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 29.

[10] Kim, H., & Mnih, A. (2018). Disentangling by Factorising. *International Conference on Machine Learning (ICML)*.

[11] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., & Lerchner, A. (2017). β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations (ICLR)*.

[12] Abdal, R., Qin, Y., & Zhu, P. (2019). Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4432-4441.

[13] Shen, Y., Zhou, P., Li, J., Ding, Z., & Loy, C. C. (2020). InterFaceGAN: Interpreting the Disentangled Latent Space of StyleGAN for Semantic Face Editing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5262-5271.

[14] Xu, Y., Yu, Y., Zhang, Y., Wu, D., Xu, Y., Fan, Y., & Ji, B. (2021). Spherical Generative Adversarial Networks. *The AAAI Conference on Artificial Intelligence (AAAI)*.

[15] Wang, F., Liu, W., Liu, X., & Liu, J. (2017). Deep Face Recognition: A Survey. *Image and Vision Computing*, 64, 84-98.

[16] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1125-1134.

[17] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2227-2235.

[18] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8789-8797.

[19] Choi, Y., Yim, J., Kim, J., Ha, J. W., & Choo, J. (2020). StarGAN v2: Diverse Image Synthesis for Multiple Domains. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1424-1433.

[20] Shih, Y. C., Lai, S. H., Huang, J. B., & Singh, M. (2019). FUNIT: Few-Shot Unsupervised Image-to-Image Translation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8667-8676.

[21] Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal Unsupervised Image-to-Image Translation. *European Conference on Computer Vision (ECCV)*.

[22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684-10695.

[23] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR)*.

[24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, 30.

[25] Bińkowski, M., Sutherland, D. J., Arbel, A., & Gretton, A. (2018). Demystifying MMD GANs. *International Conference on Learning Representations (ICLR)*.