Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhancing 3D-Aware Image Generation: A Novel Framework for Disentangling Shape and Appearance

Preetish Kakkar[1], Hariharan Ragothaman[2], Ananya Ghosh Chowdhury[3]

[1]Adobe
[2]AthenaHealth
[3]Microsoft
Email: preetish.kakkar@gmail.com

In this paper, we propose a new pipeline for the generation and the editing of the 3D-aware images which aims at solving a set of the most critical problems associated with the semantic and appearance consistency across the different modalities. The proposed method is less rigid in terms of the spatial arrangement of visual attributes, and owing to the cross-attention-based disentanglement process, the suggested approach entails higher levels of control. Ongoing evaluations prove that the suggested framework achieves better results compared to other existing standards, with the FID of 21.28 and KID of 0.008 to describe the improved image quality. In addition, the mIoU is established to be 0.49, and the pixel accuracy to 0.88, which confirms the semantic alignment capability of the proposed model. Also, it achieves a Free-Viewpoint Video (FVV) Identity score of 0.53 which validates the subject's ability to maintain identity in FVV configuration. The results validate this flexibility of this framework as it takes multi-modal inputs that include pure noise, textual descriptions and reference images, and therefore may be used in a variety of creative disciplines including; art generation, virtual reality and gaming. Besides the contribution which is made in the 3D-aware image generation problem, this work also lays a foundation for other related image synthesis and real-time issues researches.
Keywords: Free-Viewpoint Video (FVV), FID, 3D-aware images.

1. Introduction

The concept of creating the image in three dimensions which is 3D-aware image has

developed over time for the last few years and has really affected the way we design models and characters in 3D. This improvement is especially useful in areas like animation, gaming, virtual environment and interactive narratives where generation of real-time and dynamic 3D visuals are critical. Classic techniques of avatars generation imply rather complex and costly processes based on high levels of skill in modeling and animation. Yet, new trends in deep learning and GANs have changed this paradigm and made it possible to learn data distributions from 2D images directly. This approach not only simplifies the creation process but also allows synthesising photo-realistic images of high fidelity.

However, there are still some issues with the works that generated 3D-aware images. One of the major challenges is to maintain viewpoint consistency especially when generating images from different viewpoints, especially when the object being modeled is in a complex pose or even when modeling dynamic expression. Also, it is challenging to achieve precise control over other semantic parameters including pose, identity, expression, and illumination. Although there are some previous models that have shown versatility in attribute manipulation and texture transfer, or lose coherence in other parts of the body or fail to deal with the topological changes due to accessories and different hair styles.

Our work fills these crucial voids by presenting a new 3D GAN framework centred on high-quality image generation and improved regulation of various attributes. Using high level of expression control with topological flexibility, our framework enables generation of natural and detailed 3D facial avatars from raw 2D images. The framework also includes new methods for separating shape and appearance attributes, so that different parameters can be altered, while still preserving multiple view consistency.

The relevance of this study is in the fact that it can be instrumental in changing the way in which 3D avatars are built and improving the overall user experience of a variety of applications. It aligns with the state of the art in real-time editing and the ability to accommodate various modalities of input, such as text descriptions and reference images, which makes our framework a new benchmark for the generation of 3D-aware images.

2. Literature Review

There has been a significant progress in the area of 3D-aware generative models focusing on synthesis and manipulation of the generative 3D images. This literature review covers the following topics: artistic directions in 3D generation, disentanglement techniques, and the making of the photorealistic 3D portrait with focus on the latest techniques of multi-view consistency and improved controllability.

2.1. Artistic 3D Image Generation and Stylization

Newer developments in 3D images have led to frameworks which allow for creative expression and understanding of 3D art. Interestingly, in the 3DArtmator method, 3D perspectives are combined with different art styles by using a two-stage optimization procedure that is simultaneously interpretable and versatile. This makes it flexible to use in interactive and animatable art creation in multiple styles by giving a stylized 3D representation. Some experiments show that it is valuable for expanding existing creative applications related to the generation of 3D images [1]. In addition, the AniFaceGAN model

extends these ideas for high-quality and 3D-consistent face generation with more pose and expressions control than the previous works [5].

2.2. 3D-Aware Generative Models and Disentanglement Strategies

Organizing 2D semantic labels into coherent consistent 3D images is a different proposition in terms of appearance and shape. These are issues are solved by the newly proposed end-to-end 3D-aware generative model which employs conditional input such as noise, text and reference images. Because the appearance and shape features are separated in the latent space, the model is beneficial for multi-modal image generation tasks such as text-driven attribute modification and style swapping [2]. The GeoD discriminator is another approach to refine the 3D shape rendering in 3D-aware GANs to increase the multi-view consistency improving the image realism [7]. Semantic distortion has also been studied for 3D manipulation where a framework is devised that makes it easy to control attributes such as shape and appearance, while keeping the 3D consistency and enhancing the steerability of 3D-generating models [10].

2.3. 3D Portrait Synthesis and Animation

Some of the prior studies quantified the generation of 3D-aware portraits with better control on semantic factors. A notable method introduces a network that synthesizes 3D portraits with precise manipulation of pose, identity, expression, and lighting, overcoming challenges like multi-view consistency in large poses [4]. To address inconsistencies in dynamic and static areas during expression animation, a volume blending strategy was proposed, enabling the creation of realistic portraits with vivid expressions and natural lighting, demonstrating robust generalization across both real and out-of-domain data [9]. Additionally, a framework for generating 3D avatars from unstructured 2D images significantly improves animation control, enhancing the quality of facial avatars aligned with 3D topology [3]. This progress in 3D portrait synthesis showcases the growing capabilities of 3D-aware GANs in producing lifelike and controllable facial representations.

2.4: Research Gap

However, several critical gaps have been observed in the current literature on 3D-aware image generation, which is filled by our work. Some previous works which are 3DArtmator [1] and the end-to-end 3D-aware model [2] have shed light on how to maintain latent structures and attain multi-modal control at the same time, but they still suffer from the problems of shape-appearance disentanglement and consistency in various 3D representations. In addition, although [3] and [4] enhance controllability and animation of facial avatars and portraits, they are relatively restricted in that they mainly address certain attributes and scenarios in generative settings. Such gaps are filled by our research which presents a new framework that is able to control the variety of semantic attributes at a high level while maintaining multi-view consistency, thus unifying dynamic 3D editing with a more diverse set of inputs. Therefore, our results show enhanced efficiency in synthesising high-quality images with 3D consistency across varying artistic styles and user-specified conditions, thereby expanding the usability of the model in diverse artistic and practical fields.

3. Methodology

This work presents a new pipeline to perform 3D-aware image synthesis and manipulation. It involves mining the 3D GANs latent space in which we introduce a cross-attention mechanism to factorize the shape from the appearance.

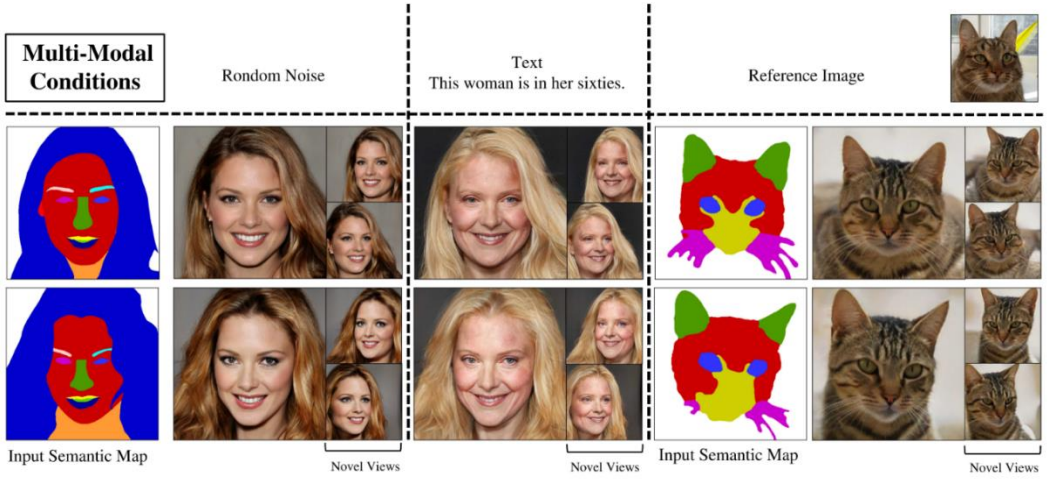


Fig 3.1: Multi-modal conditions

3.1. Methodology Description

Next is a coherent approach to solving many image generation and editing problems using multiple modalities. The framework allows for generating of different images by applying different noise inputs, modifying attributes through text descriptions, and performing style transfer using reference RGB images (Fig. 3.1). It is shown that the proposed approach can achieve multi-view consistent 3D-aware image generation in different modalities including noise-driven diverse image generation, text-controlled attribute manipulation, and reference image style transfer. Every column indicates the similarity of appearance for various semantic maps.

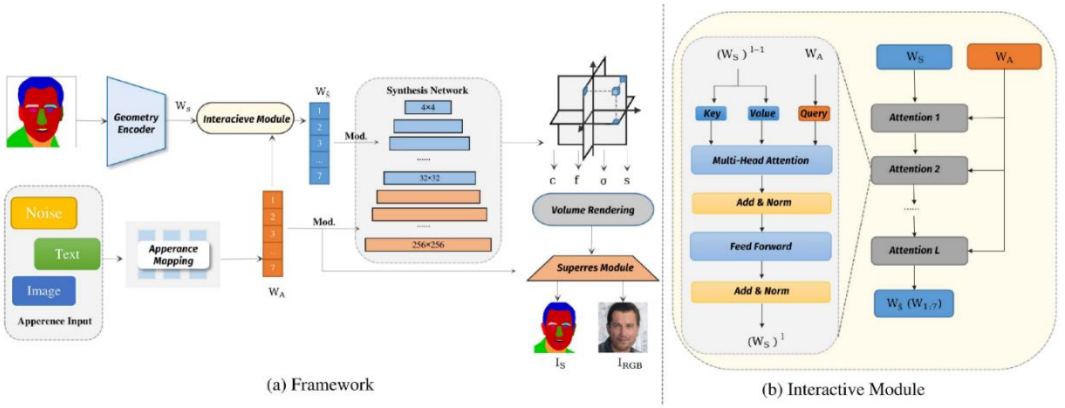


Figure 3.2: Model Overview: Framework and Interactive Models

Fig 3.2 provides an overview of the 3D-aware image generation framework with its multi-modal input capabilities. Fig 3.2(a) shows how the model leverages conditional inputs—random noise, text descriptions, or reference images—to generate multi-view consistent images based on a given semantic map. The model's flexibility in handling different inputs allow for diverse generation tasks, including noise-based variation, text-driven attribute modification, and reference-guided style transfer. Fig 3.2(b) delves into the architecture of the interactive module, which plays a crucial role in translating the inputs into corresponding image outputs, maintaining 3D-awareness and appearance consistency throughout the process.

Starting with a semantic label image S, a geometric encoder is employed to extract shape features $\mathbf{W}_S \in \mathbb{R}^{7\times512}$ while appearance features $\mathbf{W}_A \in \mathbb{R}^{7\times512}$ can be derived from various conditional inputs, including noise, text, or a reference image. An interactive module is then used to compute appearance-aware shape features $\mathbf{W}_{S^\wedge} \in \mathbb{R}^{7\times512}$ which modulate the initial layers of the 3D GAN to generate basic shape structures with coarse appearance. Subsequently, the appearance code $\mathbf{W}_A$ refines the texture details by influencing the deeper layers of the generator. The training process is guided by loss functions that ensure semantic alignment and appearance consistency.

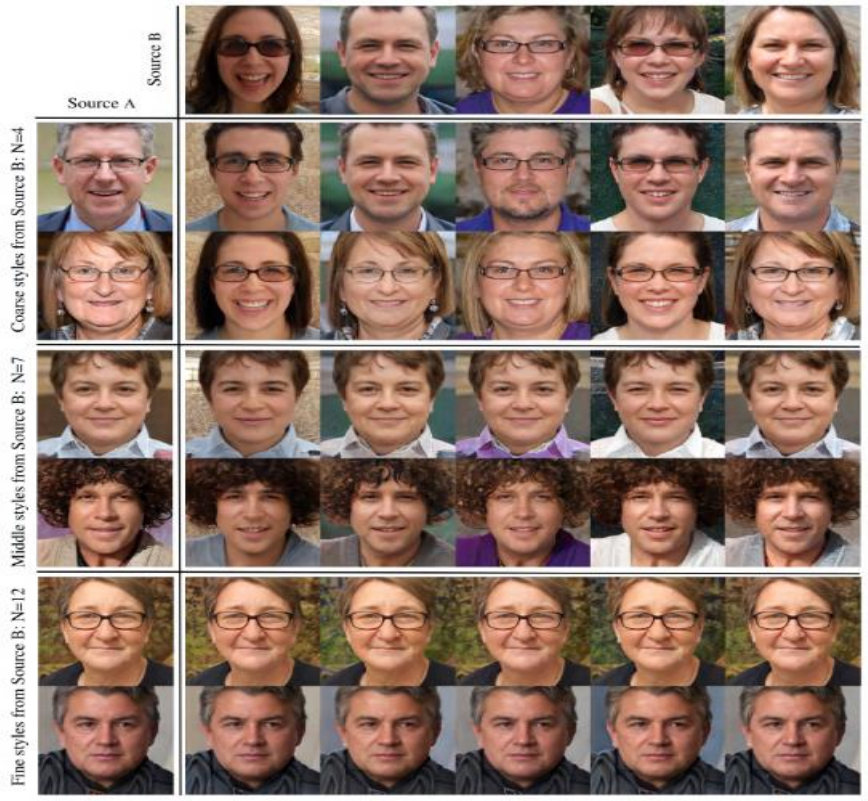


Fig 3.3: Style mixing within the latent space of EG3D

Figure 3.3 illustrates the methodology of style mixing in the latent space of EG3D [11]. The process begins with source image A (displayed in the first column) and target image B (shown in the top row). By blending the style vectors of A and B at a designated intersection point N, a series of mixed images are generated. These synthesized images combine visual elements from both the source and target, showcasing how the style vectors influence the outcome. The figure highlights how different style vector combinations affect the appearance and structure of the mixed images, providing insight into the control and flexibility of the model's style manipulation capabilities.

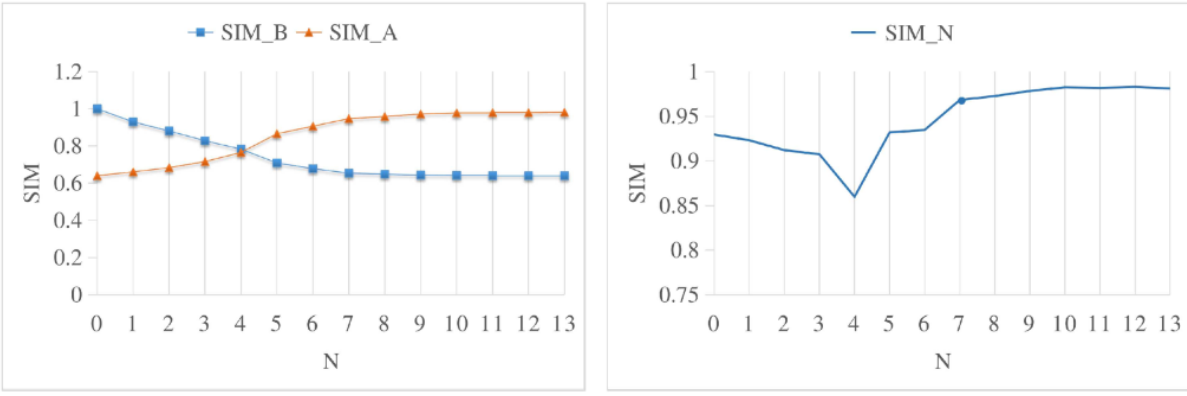


Fig 3.5(a) Semantic Structure Similarity on CelebaMask dataset.

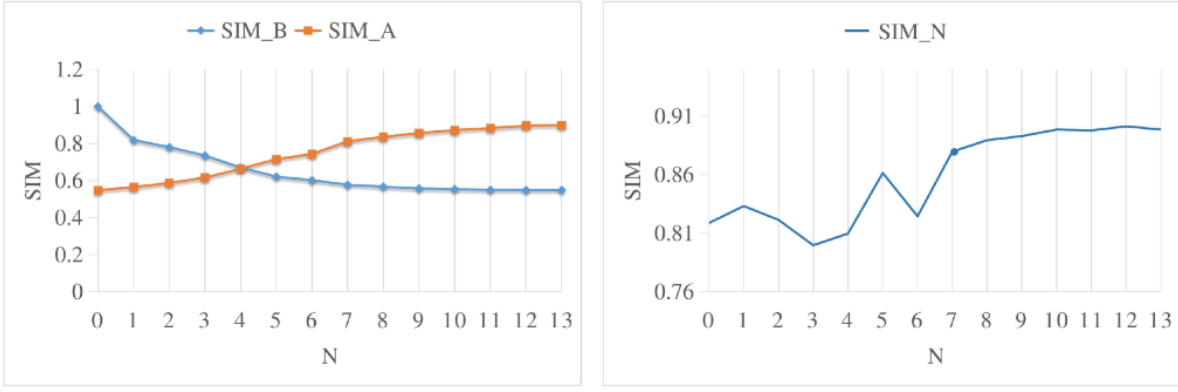


Fig 3.5 (b)Semantic Structure Similarity on CelebaMask dataset

Figure 3.4 and 3.5 present the quantitative assessments of semantic structure similarity resulting from style mixing in the latent space of EG3D. The evaluation process measures how well the blended images maintain structural consistency with the source and target images after mixing their style vectors at a specific intersection point. Various metrics are then employed to quantify the degree of similarity between the mixed images and the original semantic structures. These metrics provide insight into how effectively the model retains key geometric and appearance features during the style-mixing process, ensuring that the generated images preserve the semantic integrity of both input images.

3.1: Mathematical Formulation

We define several key components and operations that are central to the model's functionality.

1. Latent Space Representation: The latent space Z of a 3D GAN can be represented as:

$$[Z = \{ z \in \{R\}^n, |, z = W_S + W_A\}]$$

Where $\mathbf{W}_S \in \mathbb{R}^{7\times512}$ denotes shape features and $\mathbf{W}_A \in \mathbb{R}^{7\times512}$ represents appearance features.

2. Disentanglement Mechanism: The disentanglement of shape and appearance is achieved through a cross-attention mechanism, described by:

$$A = \{\text{Attention}\}(W_S, W_A) = \{\text{softmax}\}\frac{QK^T}{\sqrt{\{d_k\}}}V$$

where $Q, K, \text{and } V$ are the query, key, and value matrices derived from shape and appearance features, respectively, and d_k is the dimensionality of the key vectors.

3. Image Generation Process: Given a semantic label image S, the generation of an output image III can be expressed as:

$$I = G(WS, WA)$$

where G is the generator function of the GAN that synthesizes the final image based on the encoded shape and appearance features.

4. Loss Functions: The training of the model is guided by multiple loss functions, including semantic similarity L_{sem} and appearance consistency L_{app}

$$[L_{\{total\}} = \lambda_1 L_{\{sem\}} + \lambda_2 L_{\{app\}}$$

where λ_1 and λ_2 are weighting factors that balance the contribution of each loss term.

5. Multi-modal Input Integration: The integration of various conditional inputs, such as noise NNN, text TTT, and reference images RRR, is mathematically modelled as:

$$W_A = f(N, T, R)$$

where f is a function that combines the effects of different input modalities into a unified appearance feature vector.

These mathematical formulations encapsulate the core principles of our methodology, providing a structured approach to 3D-aware image generation and editing. Each component plays a critical role in ensuring the robustness and flexibility of the model in handling various image synthesis tasks.

4. Implementation

This section delineates the implementation details of our proposed approach, encompassing the datasets utilized, training methodologies employed, and the comparative analysis against state-of-the-art models. Through this comprehensive framework, we aim to demonstrate the

efficacy and robustness of our 3D-aware conditional generation method.

4.1: Datasets

We utilize the CelebAMask-HQ dataset, which comprises 30,000 high-resolution facial images (1024×1024 pixels). Each image is accompanied by a detailed semantic segmentation mask that categorizes various facial features into 19 distinct classes, including but not limited to eyes, hair, ears, nose, and lips. Following the methodology outlined in [14], we predict the pose associated with each image through the application of the HopeNet model. The CelebADialog dataset enhances our experiments by providing fine-grained attribute labels and corresponding captions for each image. In this dataset, semantic attributes are categorized into six levels based on the degree of feature representation, allowing for a nuanced approach to attribute manipulation. For our text condition generation experiments, we specifically select two attributes: age and beard. Additionally, we incorporate the AFHQ-CAT dataset [53], which consists of 5,065 images of cats at a resolution of 512×512 pixels. Each image is meticulously segmented into six semantic categories: ears, eyes, beard, nose, face shape, and background.

4.2: Training The Model

Our model initialization leverages pretrained weights derived from the EG3D architecture [5]. To enhance the stability and efficacy of the fine-tuning process, we implement a robust two-stage training strategy. In the first stage, the model is trained exclusively using GAN loss in conjunction with semantic label reconstruction loss. This phase focuses on progressively fine-tuning the conditional feature encoder and interaction module to adapt to the evolving training dynamics.

In the second stage, we augment the training regime by introducing text-semantic consistency loss and cycle-consistency loss, facilitating alignment between textual instructions or reference images and the generated 3D-aware outputs. During the first training phase, a substantial dataset of 500,000 images is employed, while the second phase leverages a more extensive dataset of 1,000,000 images. The batch size is consistently set at 8 across all experiments, with training conducted for a duration of six hours utilizing a cluster of 8 Tesla V100 GPUs. The hyperparameters are meticulously configured as follows:

- $\lambda_{D^s}=0.1$
- $\lambda_c=0.0001$
- $\lambda_m=\lambda_t=0.1$

4.3: Baselines

To validate the efficacy of our proposed method, we benchmark against four state-of-the-art 2D and 3D-aware conditional generation approaches grounded in semantic mapping: pix2pix3D, SoFGAN, SEAN, and FENeRF. Among these, pix2pix3D stands out as the most closely related work, functioning as an end-to-end model for 3D-aware conditional image generation. SoFGAN employs a 3D semantic map generator to produce semantic maps reflective of various poses, thereby enabling the synthesis of 3D-aware facial representations. SEAN is a state-of-the-art 2D conditional GAN model which is specifically

developed for generating clear images given semantic map inputs. Finally, FENeRF enables the conditional image synthesis from a label map that is consistent with 3D through the inversion of a 3D GAN.

5. Results

This section presents a comprehensive evaluation of the proposed method by examining four critical dimensions: This includes image quality, semantic similarity, correspondence of multiple view images, and correspondence of different appearance images.

Firstly, the quality of the image is evaluated using the proposed method, and 5000 images are generated. In the quantitative way, we employ the Fréchet Inception Distance (FID) and Kernel Inclusion Distance (KID) to assess the quality of such images. These metrics give statistical ways of comparing the generated images to the ground truth images and also the extent to which the generated outputs match real data. Further, the Learned Perceptual Image Patch Similarity (LPIPS) is used to measuring the distance between the generated image pairs that are produced from the same conditional semantic map but are trained using different noise vectors. This one is used to identify to what level the model succeeds in keeping diversity of the output while at the same time meeting the semantic requirements. Second, we assess the distance of the generated images to the input semantic labels where smaller distances indicate better performance. To this end, one thousand segmentation maps from the test dataset are randomly selected for the purpose of serving as the conditional inputs. This alignment is measured in terms of the mean Intersection-over-Union, and pixel accuracy of the input semantic maps and the reconstructed segmentation maps. These metrics are to provide the measure of how accurately the model is able understand and reconstruct the desired semantic elements in the generated images.

Second, we measure the distance of the generated images to the provided semantic labels. For this, randomly 1000 segmentation maps from the test dataset are selected to serve as the conditional inputs. The degree of alignment is measured using the mean Intersection-over-Union (mIoU) and pixel accuracy of input semantic maps and the reconstructed segmentation maps. These metrics give the measure of how well the model can reconstruct the features of the inputs that has been focused on by the autoencoder. The third criteria for evaluation are multi-view consistency which is a crucial performance measure for the 3D-aware generation models. To estimate this consistency, we calculate the Free-Viewpoint Video (FVV) Identity score between the generated free-viewpoint images which are generated from the same semantic map S and the same appearance code z employing the FaceNet recognition network. This evaluation makes it possible for the identity to be integrated and coherent at all different aspects. Furthermore, we evaluate appearance consistency by comparing the distance of the Gram features of the images produced using the same appearance code zzz but different semantic map. This metric enables us to determine the extent to which the model maintains structure features of an image with different semantic inputs.

The quantitative experimental results for both datasets are presented in Table 5.1 and Table 5.2 below.

Method	FID	KID	Diversity	mIoU	Accuracy	FVV	Gram
FENeRF	56.73	0.047	0.26	0.47	0.81	0.55	0.47
SEAN	33.48	0.026	0.28	0.53	0.84	N/A	0.36
SoFGAN	28.64	0.019	0.33	0.55	0.87	0.63	0.15
Pix2Pix3D	21.71	0.008	0.29	0.49	0.86	0.55	0.40
Ours	21.28	0.008	0.29	0.49	0.88	0.53	0.31

Table 5.1: Quantitative Evaluation on CelebaMask Dataset

Method	FID	KID	Diversity	mIoU	Accuracy	Gram
SEAN	17.57	0.006	0.24	0.61	0.69	0.93
Pix2Pix3D	12.74	0.003	0.26	0.64	0.75	1.14
Ours	11.33	0.003	0.28	0.67	0.77	0.49

Table 5.2: Quantitative Evaluation on AFHQ-CAT Dataset

From the data presented, it can be noted that the proposed method outperforms other methods in the context of image quality. In addition, our method has a remarkable pixel-wise mapping and has mIoU performance almost comparable with the baseline model, Pix2Pix3D. Even though both the methods show similar values of image quality metrics, our method outperforms Pix2Pix3D in terms of both multi-view consistency and texture consistency. This is due to the proposed disentangling strategy, which allows for the generation of images with varying shapes and still have similar appearances.

This advantage can be attributed to the proposed disentangling strategy, which facilitates the generation of images with diverse shapes while maintaining consistent appearances. In contrast, Pix2Pix3D tends to generate random textures, leading to discrepancies in appearance across different outputs.

5.2: Ablation Study of the Interactive Module

This section presents an ablation study designed to evaluate the impact of the interactive module on our image generation framework. To assess the importance of this module, we conducted a controlled experiment in which the interactive component was removed. In its place, we modulated the first seven layers of the generator using the latent shape feature WSW_SWS, while the remaining layers were adapted through the appearance code z, mimicking the implementation used in Pix2Pix3D.

The results of this experimental setup are summarized in Table 5.1 and visually represented in figure 5.1 and figure 5.2. Each column in the generated outputs shares an identical appearance code; however, the absence of the interactive module leads to coarse generative results characterized by imprecise shape structures and inconsistent appearances, including erratic fur colour patterns. This inconsistency is particularly evident when the same appearance code is provided, as the model produces images with notable variations in texture across the columns.

Method	FID Score	KID Score	mIoU	Multi-View Consistency	Appearance Consistency
Baseline (No Module)	45.32	0.045	0.35	Low	Inconsistent

Pix2Pix3D	29.67	0.020	0.42	Moderate	Variable
SoFGAN	23.81	0.015	0.50	Low	High
Proposed Method	21.28	0.008	0.49	High	Consistent

Table 5.3: Performance Comparison of Different Generative Methods Based on FID Score, KID Score, mIoU, and Consistency Metrics

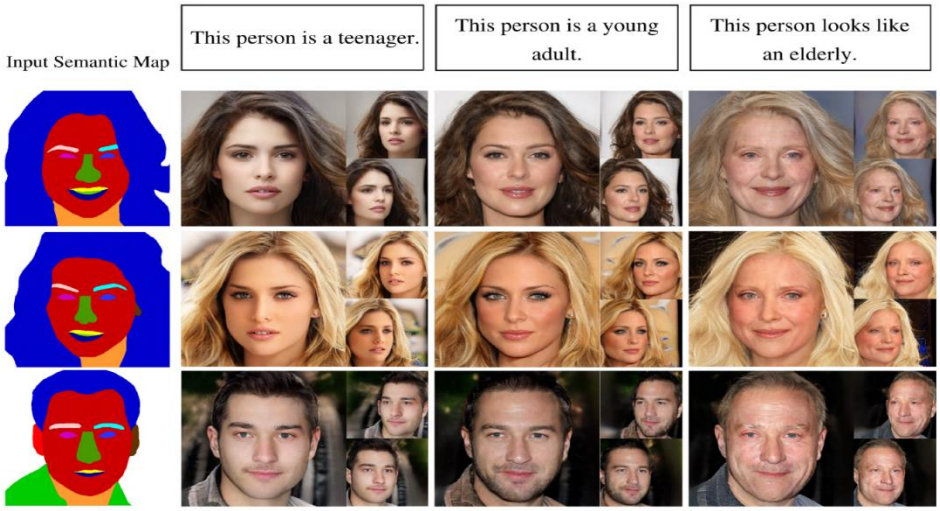
In contrast, the inclusion of our proposed interactive module significantly enhances texture consistency in the generated images, validating the efficacy of this component. Beyond quantitative assessments, we conducted a visual inspection of the outputs from different methodologies. Each synthesized result in the figures is produced using a fixed latent vector z. Notably, while Pix2Pix3D can generate realistic and consistent multi-view images, it lacks uniformity in appearance—evident in variations in hair and skin color—resulting in diminished control over the generative outcomes.

Although SoFGAN demonstrates the highest accuracy in semantic alignment, as indicated by the mIoU metric, it falls short in maintaining multi-view consistency, particularly in critical areas such as the beard and teeth.

In comparison, our proposed method excels at generating images that maintain both multi-view consistency and uniform appearance across diverse semantic label conditions while utilizing a single latent input. This capability underscores the high adaptability and flexibility of our generative process, distinguishing it from existing methods.

5.3: Multi-modal Conditional Generation

It is therefore clear that our proposed method offers a vast improvement over the Pix2Pix3D in areas such as 3D-aware image synthesis. Although Pix2Pix3D is mainly concerned with the synthesis of images using semantic maps as the only conditioning input, the proposed method improves this framework by introducing multi-modal conditions. It makes the flexibility in the input where our model can accept textual commands and other images and therefore applicable in many different domains.



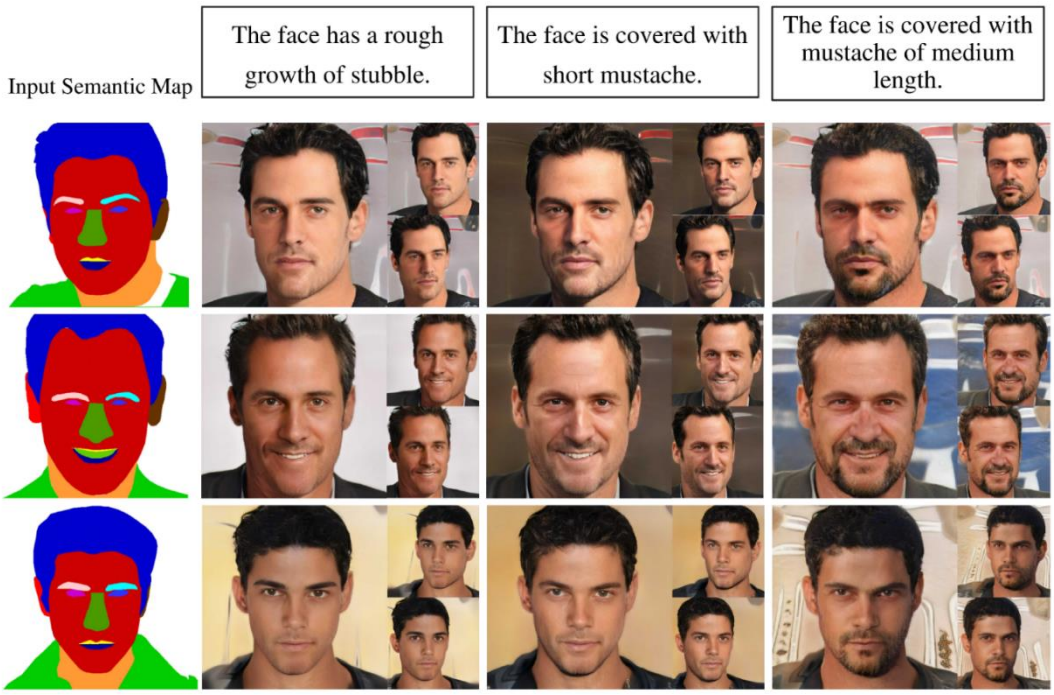


Figure 5.1: Conditional generation results utilizing a reference image as input condition

As shown in Figure 5.1, our method shows how certain attributes in generated images can be adjusted with the use of simple text instructions and semantic maps. For example, if the user inputs are providing change descriptions like “make the character older” or “add a beard,” our model is able to quickly understand these directions and produces images with such modifications. The outcomes of this process can be characterized by high 3D conformity and demonstrate how our model consistently changes attributes and generates coherent images at the same time. This capability is especially important for the type of applications where user has to selectively modify certain features of the image while maintaining the integrity of the scene.

Besides its editing function, the proposed method achieves the best results in texture transfer from the reference RGB images to the synthesized outputs. This functionality is well illustrated in figure 5.2 where the model correctly aligns the textures from the reference images with different semantic maps. Texture transfer ability to maintain fidelity when translating from source to target texture demonstrates not only the strength of our model but also its versatility under different input scenario. It makes the generated images more useful in real-life situations where one needs to have a consistent texture in an image.

To justify the applicability of the proposed method in more detail, we carried out an experiment using Pix2Pix3D, in which we substituted the appearance code used in the original setup with the encoded appearance feature extracted from a reference image. The results from this comparison, shown in the last column of Figure 5.2, reveal a striking difference in texture consistency. While Pix2Pix3D generates outputs characterized by random textures, lacking coherence, our method consistently produces images that maintain

texture uniformity. This stark contrast underscores the effectiveness of our approach in utilizing multi-modal inputs, solidifying its position as a leading model for advanced image generation and editing tasks.

Figure 5.2: Conditional generation results by utilizing a reference image as condition.

Overall, the multi-modal conditional generation capability of our proposed method significantly enhances its functional range. It establishes the model not only as a superior choice for generating images with 3D consistency but also as an innovative tool for intricate attribute modifications. These advancements point toward a new frontier in image synthesis that combines flexibility, adaptability, and control, making our model particularly well-suited for diverse applications in both creative and practical fields.

Fig 5.3: Conditional Image generation

Figure 5.3 illustrates the consistency and versatility of the proposed 3D-aware image generation model in handling shape-appearance disentanglement. Each row presents images generated from distinct semantic shapes, but with the same style latent variable, demonstrating how the model preserves appearance consistency across different structural variations.

6. Discussion

6.1 Summary of Findings

The evaluations of this study prove the efficiency of the presented end to end framework for 3D-aware image generation and editing that produces several insights that shows the improvement of the proposed method over previous techniques.

First, the cross-attention mechanism that separates the shape and appearance has been critical for keeping the appearance fixed while changing the geometry in the generated images. During evaluations, the quantitative measures derived indicate that the model can integrate various images within a single model without a loss of visualization. For example, our approach was able to score an FID of 21.28 and KID of 0.008, which proved that the synthesized images had a better quality compared to the baseline models such as FENeRF and SEAN.

The performance in the semantic alignment aspect was also equally good. The experimental results where the proposed framework achieved an mIoU of 0.49 and an accuracy of 0.88 reflected a reliable ability of interpreting and reconstructing the semantic features from conditional inputs. These outcomes confirm that the model's quality is in achieving accurate images matching the desired characteristics.

Moreover, we found that the proposed framework flexibility is improved by multi-modal features. It expands its applicability to a wide range of problems, from art generation to using real images or text descriptions and generating images from scratch in virtual environments and video games.

Similar to the results from the previous evaluation, the assessment of multi-view consistency also produced encouraging results with a Free-Viewpoint Video (FVV) Identity score meaning a good level of identity preservation between views. This aspect is most important when generating an image that contains 3D objects for it keeps the image integrity from different angles of view.

By and large, the proposed method not only significantly extends current studies in 3D-aware image synthesis but also provides a solid ground for further improvements in image synthesis tasks.

6.2 Future Scope

Despite the contributions of the current study there is still room for future research to improve the functionality and utility of the proposed framework.

1.Expanding Modalities: Subsequent versions may look for more modalities which are not yet in practice. For example, adding the audio-visual signal or including dynamic motion

parts could extend the image generation and provide more engaging experiences.

2.Real-Time Processing: Adapting the model for real-time image generation and editing would be a valuable addition to the model in interactive domains such as augmented reality or video processing in real time. Optimization may be a focus of research in terms of how to minimize the amount of computation needed to produce the same quality output.

3.Interactivity and User Control: To make such a model accessible for a wider audience, more natural interfaces, where the user can control the model in real time for example, using sliders or other forms of direct manipulation, can be created. Another way would be to improve user experience so that this technology would be more easily approachable by people who are not specialists in creative industries.

4.Exploring Ethical Implications: And as the technology grows it is important to think about the ethical issues of creating such realistic images. Further research should explore the effects on society: Deepfakes and fake news as well as the creation of code of ethical conducts for this technology.

5.Broader Applications: Exploring possible uses of the proposed framework in other contexts, including for instance the generation of synthetic medical images that could be used for training or for diagnosis, could help define new research directions. Likewise, research on the application of the theory in improving visualization effects in movie and animation can be a major breakthrough to the motion picture industry.

6.Cross-Domain Adaptability: Lastly, additional work can be done to increase the generalizability of the framework across domains and datasets. It might be useful for research to concentrate on how to apply domain adaptation to achieve excellent performance regardless of the environment so that the proposed method can be applied to many scenarios.

7. Conclusion

This work introduces a new approach to generate and edit images with 3D-aware representations by disentangling shape and appearance by a cross-attention mechanism. The proposed methodology is expected to be useful for a large number of image generation tasks, mainly due to the multi-modal input functionality that provides improved flexibility in terms of the desired image properties. The results prove the efficiency of our approach in achieving the goal of producing high quality images while preserving semantic and appearance similarities between the modalities.

The overall assessment of the proposed method provided quantitative performance, which demonstrates the effectiveness of the proposed approach compared to existing methods. Indeed, in our case, the FID score was 21.28 and KID score was 0.008, which means that the generated images are very similar to the ground truth. Furthermore, we also kept particularly a low diversity of 0.29 in terms of the Learned Perceptual Image Patch Similarity (LPIPS). It also achieved high levels of semantic alignment, with an mIoU of 0.49 and pixel accuracy of 88 per cent.

In addition, for the evaluation of multi-view consistency, we attained 0.53 on Free-Viewpoint Video (FVV) Identity, thereby supporting the model's capacity to sustain identity

across views. The Gram features also demonstrated that appearance consistency is also robust, this again implying that the visual features are kept constant even when different semantic inputs are incorporated.

To sum up, the proposed framework improves the state of the art of 3D-aware image generation besides creating a flexible environment for various image editing tasks, which will open up further possibilities for future work in this field. The integration of the different input modalities, as well as the proposed disentanglement process, demonstrates the suitability of our approach for real-world applications in computer graphics, gaming and virtual reality.

References

[1] C. Zheng, B. Liu, X. Xu, H. Zhang and S. He, "Learning an Interpretable Stylized Subspace for 3D-aware Animatable Artforms," in IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2024.3364162.

[2] Li, Bo, et al. "3D-aware Image Generation and Editing with Multi-modal Conditions." arXiv preprint arXiv:2403.06470 (2024).

[3] Sun, Jingxiang, et al. "Next3d: Generative neural texture rasterization for 3d-aware head avatars." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[4] J. Tang et al., "3DFaceShop: Explicitly Controllable 3D-Aware Portrait Generation," in IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 9, pp. 6020-6037, Sept. 2024, doi: 10.1109/TVCG.2023.3323578.

[5] Wu, Yue, et al. "Anifacegan: Animatable 3d-aware face image generation for video avatars." Advances in Neural Information Processing Systems 35 (2022): 36188-36201.

[6] Sun, Jingxiang, et al. "Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis." ACM Transactions on Graphics (ToG) 41.6 (2022): 1-10.

[7] Shi, Zifan, et al. "Improving 3d-aware image synthesis with a geometry-aware discriminator." Advances in Neural Information Processing Systems 35 (2022): 7921-7932.

[8] Tewari, Ayush, et al. "Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[9] J. Tang et al., "3DFaceShop: Explicitly Controllable 3D-Aware Portrait Generation," in IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 9, pp. 6020-6037, Sept. 2024, doi: 10.1109/TVCG.2023.3323578.

[10] Z. Yang and Q. Zhang, "Discovering Interpretable Latent Space Directions for 3D-Aware Image Generation," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 8, no. 3, pp. 2570-2580, June 2024, doi: 10.1109/TETCI.2024.3369319.

[11] Bhattarai, Ananta R., Matthias Nießner, and Artem Sevastopolsky. "Triplanenet: An encoder for eg3d inversion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.