NeRF-Diffusion for 3D-Consistent Face Generation and Editing

Héctor Laria[1,2], Kai Wang[1], Joost van de Weijer[1,2], Bogdan Raducanu[1,2] and Yaxing Wang[3]
[1]*Computer Vision Center, Barcelona, Spain*
[2]*Universitat Autònoma de Barcelona, Spain*
[3]*Nankai University, China*

Keywords: NeRF, Diffusion Models, 3D Generation, Multi-View Consistency, Face Generation.

Abstract: Generating high-fidelity 3D-aware images without 3D supervision is a valuable capability in various applications. Current methods based on NeRF features, SDF information, or triplane features have limited variation after training. To address this, we propose a novel approach that combines pretrained models for shape and content generation. Our method leverages a pretrained Neural Radiance Field as a shape prior and a diffusion model for content generation. By conditioning the diffusion model with 3D features, we enhance its ability to generate novel views with 3D awareness. We introduce a consistency token shared between the NeRF module and the diffusion model to maintain 3D consistency during sampling. Moreover, our framework allows for text editing of 3D-aware image generation, enabling users to modify the style over 3D views while preserving semantic content. Our contributions include incorporating 3D awareness into a text-to-image model, addressing identity consistency in 3D view synthesis, and enabling text editing of 3D-aware image generation. We provide detailed explanations, including the shape prior based on the NeRF model and the content generation process using the diffusion model. We also discuss challenges such as shape consistency and sampling saturation. Experimental results demonstrate the effectiveness and visual quality of our approach.

1 INTRODUCTION

Producing high-fidelity images in a 3D-consistent manner while simultaneously capturing the geometry of objects from image collections without the need for any 3D supervision is referred to as 3D-aware generative image synthesis (Xia and Xue, 2022). This capability is valuable in various applications, including virtual reality, robotics, and content creation. While current methods excel at 3D shape estimation using NeRF features (Gu et al., 2022), SDF information (Or-El et al., 2022), or triplane features (Chan et al., 2022), they fall short in guiding the generation process based on specific prompts or other conditional factors once the model is trained.

On the other hand, recent advancements in text-to-image (T2I) generators (Rombach et al., 2022; Ramesh et al., 2021; Ramesh et al., 2022; Saharia et al., 2022; Midjourney.com, 2022) have allowed for realistic image synthesis based on widely different textual description, but even these models have limitations when it comes to embracing novel conditions. To overcome these limitations, various solutions have been proposed, such as ControlNet (Zhang and Agrawala, 2023), T2I-adapter (Mou et al., 2023), GLIGEN (Li et al., 2023b), Universal Guidance (Bansal et al., 2023), which enable T2I models to accept additional conditions like sketches, segmentation maps, bounding boxes, and layout images. However, these approaches do not address the problem of 2D novel view synthesis. A naive solution is to condition the Stable Diffusion model with 3D features extracted from 3D-aware image synthesis methods, as shown in Fig. 2 (top). However, this approach does not guarantee identity preservation. Another issue is how to extend T2I models with 3D capabilities. Several methods (Poole et al., 2022; Seo et al., 2023; Tang et al., 2023) have successfully achieved novel views of a target, but these are constrained to learning a model for a single subject.

In this paper, we present a novel approach for generating 3D-aware image synthesis with text-editing capability. Our method breaks down this task into simpler sub-tasks and leverages off-the-shelf pretrained architectures for each stage. By doing so, we take advantage of the rich representations learned by these pretrained models, requiring only modest resources to achieve convincing results. The proposed

Laria, H., Wang, K., van de Weijer, J., Raducanu, B. and Wang, Y.
NeRF-Diffusion for 3D-Consistent Face Generation and Editing.
DOI: 10.5220/0012381500003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 587-594
ISBN: 978-989-758-679-8; ISSN: 2184-4321

approach consists of two main components: a shape prior based on a pretrained Neural Radiance Field (NeRF) model and a diffusion model for content generation. The NeRF model provides a robust shape representation that is consistent among viewpoint translations. We condition the NeRF model with a shared style vector to control the scene's identity. The diffusion model is based on Stable Diffusion (SD), a pretrained leading T2I generator with outstanding quality and T2I capabilities. By conditioning the SD model with 3D features extracted from 3D-aware image synthesis methods, we enhance its capability to generate novel views with 3D awareness. However, maintaining 3D consistency during the iterative sampling process of the diffusion model is challenging. To address this, we introduce a consistency token that is shared between the NeRF module and the diffusion model, ensuring cohesion between style and shape. An additional advantage of our framework is its ability to replace the consistency token to modify the style among 3D views using text editing. This feature allows users to manipulate and modify the generated views while preserving their inherent semantic content. We showcase these capabilities in Figure 1 (middle, bottom).

Our contributions include:

- Successfully incorporating 3D awareness into a T2I model,
- Addressing the challenge of identity consistency in 3D view synthesis and
- Enabling text editing of 3D-aware image generation.

2 RELATED WORK

3D Generative Models for Novel View Synthesis from Single-View Images. The aim of 3D-aware generative image synthesis is to produce high-quality images with 3D consistency, capturing object surface details from 2D image collections without explicit 3D supervision. Early works, like GRAF (Schwarz et al., 2020) and its successors, leverage NeRF and adversarial frameworks for realistic scene representation and training. CIPS-3D (Zhou et al., 2021) combines shallow NeRF and deep 2D Implicit Neural Representation (INR) for shape and appearance synthesis, while StyleSDF (Or-El et al., 2022) integrates SDF-based 3D representation into StyleGAN. EG3D (Chan et al., 2022) proposes a triplane hybrid 3D representation, employing a StyleGAN2 generator for image rendering with super-resolution. StyleNeRF (Gu et al., 2022) enhances rendering efficiency by integrating NeRF into a style-based generator for high-resolution 3D image generation.

However, these models are often domain-specific and lack generalization to unseen datasets. To overcome this limitation, we propose a method based on StyleNeRF, a state-of-the-art 3D GAN. Our approach enables novel 3D-aware view synthesis from a single view through fine-tuning a pretrained Text-to-Image (T2I) model, requiring only a single finetuning iteration.

3D-Aware Text-Edit Image Synthesis. DreamFusion (Poole et al., 2022) pioneered Score Distillation Sampling (SDS), utilizing a frozen diffusion model as a critic for NeRF model learning. SDS allowed users to specify subjects through text prompts, but it was limited to one subject at a time during training. 3DFuse (Seo et al., 2023) addressed prompt ambiguity by optimizing an initial image and introduced a semantic identity space for consistency tokens. Make-it-3D (Tang et al., 2023) learned a NeRF model and a texture point cloud in two stages for single-object synthesis. In contrast, our method exploits diffusion model-generated views for high-quality generation without multi-stage processing. All these methods distill knowledge from a diffusion model, while our approach uses pretrained NeRF models as geometric priors, employing the diffusion model for generating consistent views across diverse subjects. Importantly, our method is not limited to single-subject generation.

T2I Model Augmentation with Additional Conditions. To enrich Text-to-Image (T2I) models with fine-grained details, studies extend T2I diffusion models with additional conditioning modalities. Composer (Huang et al., 2023) explores multiple control signals alongside text descriptions, training from scratch with extensive datasets but incurring high GPU and financial costs. Leveraging pretrained T2I models like Stable Diffusion (Rombach et al., 2022), ControlNet (Zhang and Agrawala, 2023) integrates various controls into the denoising U-Net, enabling task-specific guidance. GLIGEN (Li et al., 2023b), T2I-Adapter (Mou et al., 2023), and others use large-scale pretrained T2I diffusion models for customized image generation, partially fine-tuning them. Some methods (Mou et al., 2023; Hu et al., 2023) offer composable controls for robust augmentation. To address this, we enhance ControlNet for 3D-aware image synthesis, maintaining identity consistency with depth or NeRF features. We introduce a module to improve handling of real image inputs and obtain corresponding 3D views.

Editing Methods for NeRF with Prompts. Simplifying the diffusion generation to an optimization-based prompt-guided approach draws parallels with

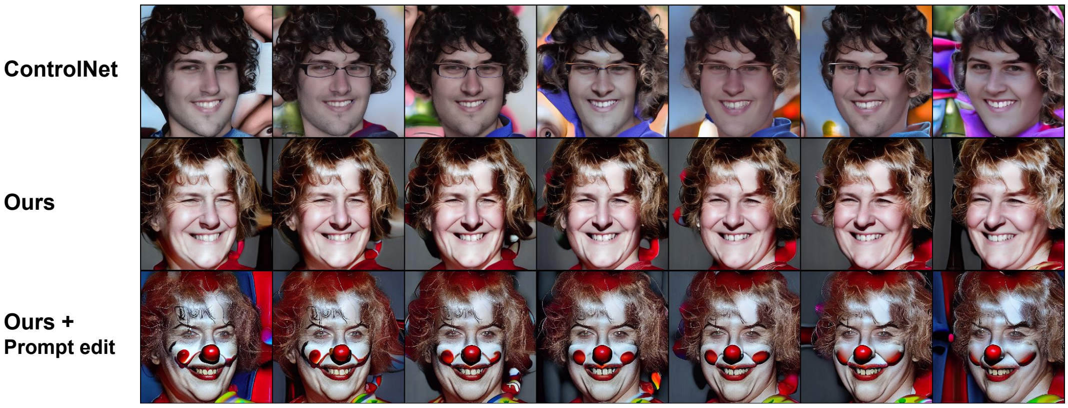

Figure 1: Naive generation of 3D-aware content with T2I models (top). Our approach links the identity among views (middle) and is amenable to consistent editing (bottom). This figure shows that ControlNet can obtain high quality visual per frame results, however the generated images lack 3D consistency as can be seen by the presence and absence of glasses for some viewpoints, and inconsistent background.

recent studies, notably StyleNeRF (Gu et al., 2022). This involves latent code optimization using a CLIP loss guided by a text prompt. CLIP-NeRF (Wang et al., 2022) disentangles a pretrained NeRF, translating the CLIP latent space to condition the NeRF model. NeRF-Art (Wang et al., 2023) modulates a pretrained NeRF model for shape and appearance through a perceptual loss mechanism with CLIP. Instruct-NeRF2NeRF (Haque et al., 2023) employs an external diffusion model for data modification akin to knowledge distillation. Instruct 3D-to-3D (Kamata et al., 2023) adjusts the NeRF model directly, reminiscent of DreamFusion (Poole et al., 2022), using score distillation. These methods require fine-tuning or supplementary modules for each prompt-guided edit. In contrast, our approach utilizes generic NeRF features, adjusting appearance seamlessly without retraining or specific adjustments for each edit. Zero-1-to-3 (Liu et al., 2023) is closely related but requires extensive fine-tuning of the entire diffusion model and dataset compilation with camera extrinsics. Our methodology achieves satisfactory outcomes without relying on data, employing a generic explicit shape model and only requiring a control module for shape feature interpretation and diffusion guidance.

3 METHODS

Generating 3D-aware image synthesis is a challenging task that requires inferring and preserving both geometry and textures. In this paper, we present a novel approach that breaks down this complex task into sub-tasks and utilizes off-the-shelf pretrained architectures for each sub-task. Our approach enables us to achieve high-quality results by leveraging the rich representation of these pretrained models, requiring modest additional training.

Firstly, we employ a pretrained neural radiance field model (NeRF) to estimate a robust shape prior that is consistent among viewpoint translations. This prior conditions the pretrained diffusion model that generates the final image. However, due to the iterative nature of diffusion model sampling, changes in the conditioning result in changes in the output, leading to a loss of 3D consistency. Our method solves this by introducing a consistency token that automatically maintains the desired details during the generation of new views. This token is shared between the NeRF module and the diffusion model to achieve cohesion between style and shape.

The proposed framework, illustrated in Figure 2, leverages prior knowledge of several separated models to learn a new task in a zero-shot fashion. This approach enables additional capabilities like image-to-3D inversion, zero-shot image translation, and editing as shown in Figure 1. In the following sections, we introduce the method and its components in detail.

3.1 Shape Prior

We employ a pretrained Neural Radiance Field (NeRF) representation (Mildenhall et al., 2020), which has shown state-of-the-art results in 3D-scene reconstruction and novel view synthesis tasks. NeRFs are typically parametrized by multi-layer perceptrons (MLPs) $f : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ the scene, where $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{d} \in \mathbb{R}^2$ represent the position and viewing direction, respectively. The output of the MLPs includes a volume density $\sigma \in \mathbb{R}^+$ and a view-dependent color $\mathbf{c} \in \mathbb{R}^3$.

To ensure representation coherence between the

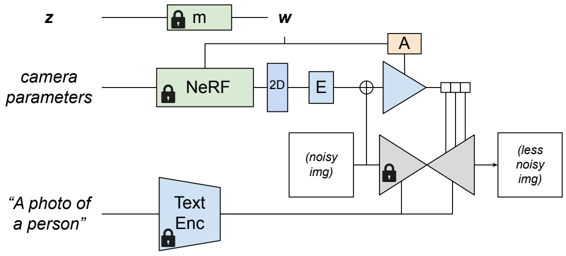

Figure 2: The proposed method consists of a controllable diffusion model augmented with a spatial prior and a consistency token for consistent geometry and identity preservation among different views.

shape and texture of the generated scenes, we condition the NeRF model with a shared style vector $\mathbf{w}$ (Gu et al., 2022; Karras et al., 2020) that controls the person's identity. The conditioned module is denoted by $f_{\mathbf{w}} : (\mathbf{x},\mathbf{d}) \rightarrow (\mathbf{c_w},\sigma_{\mathbf{w}})$, where $\mathbf{w} = m(\mathbf{z}), \mathbf{z} \in \mathcal{Z}$. Following the StyleGAN (Gu et al., 2022) architecture, $m(\cdot)$ is a mapping network and $\mathcal{Z}$ is the unit Gaussian sphere where noise is drawn from.

We perform 2D aggregation on the output of the NeRF model to reduce the dimensionality of the conditioning map. This enables efficient computation of the NeRF output and reduces the number of parameters required for the final model. The modified conditioning map is used to generate a view of the 3D scene through a subsequent diffusion model.

However, rendering high-resolution images using NeRF is computationally demanding due to its pixel-wise ray computation. Conversely, low-resolution renders are more efficient but may compromise 3D consistency when performing pixel-space operations like upsampling. To find a balance between efficiency and consistency, we sample at the same spatial resolution as the latent features $\mathbf{z} \in \mathbb{R}^{C\times 64\times 64}$ of the diffusion model. Subsequently, as explained in the next section, the encoder $E(\cdot)$ of the next stage is modified to ensure spatial consistency.

3.2 Content Generation

There are several image generation models that can incorporate spatial conditioning, such as GANs, VAEs, and NeRFs. In this study, we choose diffusion models due to their high quality, output diversity, and flexible inference conditioning, including their ability to allow for text-based image editing.

One way to train a conditioned generation model is by means of a ControlNet (Zhang and Agrawala, 2023) approach, which learns a zero-convolution gated module to interact with the decoder of the frozen generative model. However, we found that a naive application of ControlNet to this problem suffers from *identity drift*, referring to the fact that different view conditions modify the denoising path. This effect leads to a significantly different output even with the same initial conditions, as seen in Figure 1 (top). Our hypothesis is that this occurs because, at each step, the diffusion model lacks constraints within the space of possible textures to approximate. Consequently, it tends to derive general (and thus inconsistent) guidelines from both the current denoising state $\mathbf{x}_t$ and the text conditioning.

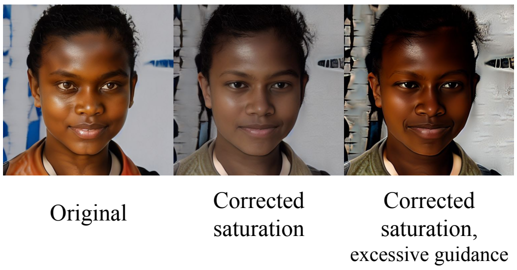

Figure 3: Comparison of sample quality. Identity is correctly maintained even on saturated conditioning (right).

To secure the identity among random states, we propose learning an injective translator function $f : \mathcal{W} \rightarrow \mathcal{T}$ that maps any style-space instance in $\mathcal{W}$ to the token space $\mathcal{T}$. This enables us to condition the diffusion model with concrete style, similar to how text would condition the model. This translator is implemented using an adapter network $\mathbf{t_w} = A(\mathbf{w})$ translates the style vector $\mathbf{w}$ to the corresponding textual token $\mathbf{t_w}$, and both the conditioning modules and adapter network are jointly learned. We call $\mathbf{t_w}$ an *identity token* throughout the text, as it comprises all the condensed information of the current style and can act as a powerful word embedding. It is so that the parametrization of $A(\cdot)$ by a linear layer is enough to harness its conditioning. Our final architecture (Figure 2) produces significantly improved results, as shown in Figure 1 (middle). To avoid interfering with the text conditioning capabilities of the diffusion model, we repurpose some unused text tokens to input this information into the network cross-attention mechanism, a common practice in other works (Gal et al., 2022; Han et al., 2023; Li et al., 2023a).

To ensure shape consistency, we modify the input encoder $E(\cdot)$ of ControlNet by replacing the upsampling and following downsampling blocks with 1×1 convolutions to avoid pixel-space operations that may cause spatial inconsistency.

3.3 Sampling Saturation

In our proposed method, we employ a conditioning module during the sampling process to guide the generation of images with a specific shape, and we utilize

Table 1: Method ablation of all the proposed contributions. TL1, TL2, TL4 are view consistency metrics. FI is an identity consistency metric. FID is an image quality metric.

Method	TL1 ↓	TL2 ↓	TL4 ↓	FI ↓	FID ↓
ControlNet	0.1875	0.1966	0.2160	0.2852±0.13	29.76
+ identity (saturated)	0.1489	0.1717	0.2075	0.0840±0.03	38.33
+ identity	0.1332	0.1545	0.1786	0.1188±0.05	32.67

classifier-free guidance (Ho and Salimans, 2021) for texture and view-conditional details such as light reflections and shadows. However, we have observed that the use of classifier-free guidance can result in style saturation, leading to excessive values for color, lighting, and other accents at any guidance scale. We call this phenomenom *sampling saturation*. A similar result can be experienced when the guidance scale is excessively tuned up, as shown in Figure 3.

Concretely, in order to overcome this undesirable effect, our method extends the text conditioning of N token embeddings to include an identity token as $y = (t_{1,\cdots,N}, t_{\text{id}})$, respectively. The conditioning is sampled as usual,

$$\nabla_x \log p_\gamma(x \mid y) = \nabla_x \log p(x) + \gamma(\nabla_x \log p(x \mid y) - \nabla_x \log p(x)), \quad (1)$$

where γ is the guidance scale. Let us consider the conditioned model $p^{\text{CN}}(x)$ that adds residual information to the frozen generative model, effectively modifying the output distribution by $p(x)r(x \mid c)$, being $p(x)$ the previous knowledge and $r(x \mid c)$ the learned residual control given the conditioning input c. If we apply guidance sampling to this model, it yields

$$\begin{aligned}\nabla_x \log p_\gamma^{\text{CN}}(x \mid y, c) &= \nabla_x \log p^{\text{CN}}(x) + \\ &\quad + \gamma(\nabla_x \log p^{\text{CN}}(x \mid y, c) - \nabla_x \log p^{\text{CN}}(x)) \\ &= \nabla_x \log p(x) + \nabla_x \log r(x \mid c) + \\ &\quad + \gamma(\nabla_x \log p(x \mid y) + \nabla_x \log r(x \mid y, c) - \\ &\quad - \nabla_x \log p(x) + \nabla_x \log r(x \mid c)). \quad (2)\end{aligned}$$

We observe that the unconditional control term $\nabla_x \log r(x \mid c)$ lacks the identity consistency information embedded in y, which causes it to push the evaluation away from the desired identity t_{id} at each step. To overcome this mismatch, we introduce the identity token back into the unconditional term for aligned generation, which leads to the following equation used for sampling:

$$\begin{aligned}&\nabla_x \log p_\gamma^{\text{CN}}(x \mid t_{1,\cdots,N}, t_{\text{id}}, c) = \\ &\quad \nabla_x \log p(x) + \nabla_x \log r(x \mid c, t_{\text{id}}) + \\ &\quad + \gamma(\nabla_x \log p(x \mid t_{1,\cdots,N}, t_{\text{id}}) + \\ &\quad + \nabla_x \log r(x \mid t_{1,\cdots,N}, t_{\text{id}}, c) - \\ &\quad - \nabla_x \log p(x) + \nabla_x \log r(x \mid c, t_{\text{id}})). \quad (3)\end{aligned}$$

By including the identity token in the unconditional term of the guidance, we ensure that the model is aware of the true identity of the object being generated, even if the learned conditioning is heavily biased towards other text conditioning. This prevents the accumulation of identity error during iterative sampling and leads to more stable and accurate image generation. This adjustment can be extended to any work that uses this type of conditioning on pretrained models and incorporates information not present in the spatial conditioning input c.

3.4 Training

During training, we first sample a tuple of 2D aggregated NeRF features $\mathbf{c}$, style vector $\mathbf{w}$ and generated image $\mathbf{x}$ from a pretrained frozen StyleNeRF model. The features are passed to the ControlNet module, along with the conditioning text and style (now converted to identity token by the adapter), to produce residuals for the frozen diffusion model. The frozen diffusion model receives a noised version of the image $\mathbf{x}$, the conditioning text and the residuals produced, to generate an ε output following ε-parametrization from Stable Diffusion. The whole architecture is driven by the L2 distance between the estimated and added noise at that denoising step.

It's important to highlight that, despite 3D-aware generative models like StyleNeRF and EG3D lacking text-editing capabilities compared to current diffusion model methods, we can inherit the text-editing property of the original text-to-image diffusion model in Stable Diffusion by conditioning it with NeRF features. Importantly, this inheritance occurs without requiring additional training or model augmentations.

4 EXPERIMENTS

4.1 Experimental Setup

Dataset. For training, we use a pretrained StyleNeRF on FFHQ (Karras et al., 2021) in a zero-shot fashion, simply using the generated output resized to 512 pixel resolution.

Evaluation metrics. We assess our method using a temporal loss (TL) (Wang et al., 2020) to mea-

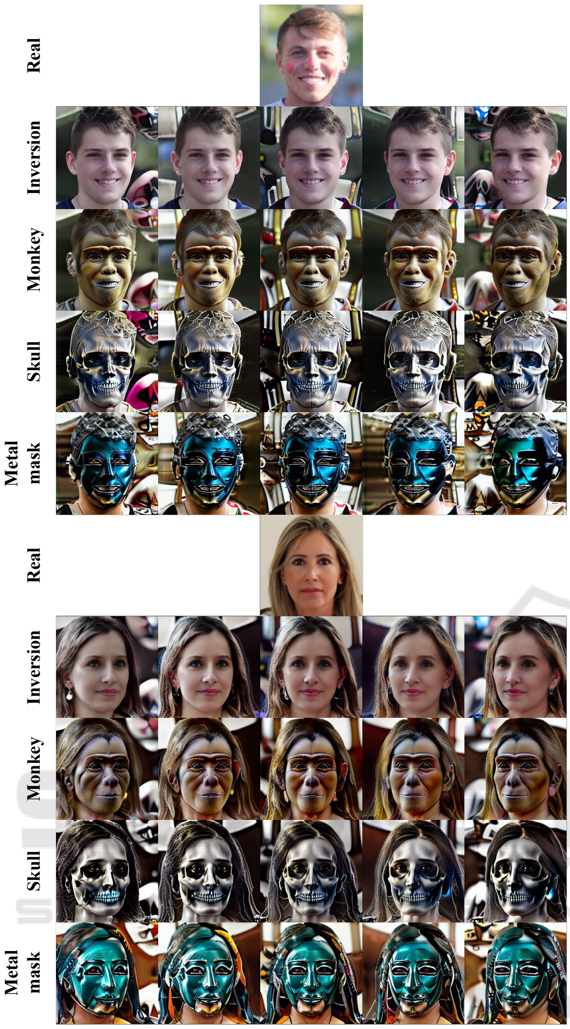

Figure 4: Generation of novel views given a real image. Generation of 3D-aware content base on text editing of the original subject.

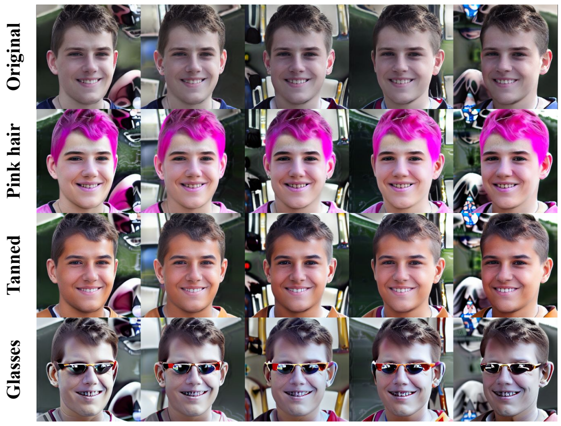

Figure 5: Further generation of 3D-aware content base on text editing.

sure scene consistency, since the problem we are facing in evaluating the quality of our outputs is similar to those found in video generation problems. It is a crucial metric to ensure that the generated sequences exhibit smooth transitions, maintain realism over time, and align with human perception. To further evaluate our method, we make a sequence where we have subject and background and only vary the viewing angle. To make sure the identity is consistent, we employ one more face-specific metric we called *Face Identity* (FI), by comparing face embeddings from a VGG-Face network of different views to a reference front-facing generation. We use the LightFace (Serengil and Ozpinar, 2020) framework for this metric, by computing the mean of all distances of all views to the reference image. We also show the standard deviation for completeness. Additionally, visual quality is also measured to make sure the approach retains the same or similar quality as the base models, although this measure does not take 3D consistency into acount. We use the well-known metrics Fréchet Inception Distance (FID) to measure sets of single frames versus real data.

4.2 Results

Ablation. In our experiments, we made several design choices to evaluate the performance and effectiveness of our proposed method. Firstly, we utilized the NeRF++ model (Zhang et al., 2020) as our base neural radiance field model. This choice was motivated by the model's ability to disentangle foreground and background. Additionally, the NeRF++ model has been pretrained on faces (Gu et al., 2022), making it suitable for our face generation task.

To provide a baseline for comparison, we employ ControlNet and evaluate its performance. We perform a series of experiments to extend it to achieve consistent identity generation, as shown in Table 1.There we showcase enhanced view consistency across consecutive frames, namely, the next, second, and fourth frames (TL1,2,4). We also explore subject identity consistency (FI). We attribute this improvement to the improved encoder that was able to extract finer details from the object and the background, as well as the introduction of the consistency (or identity) token for controllable diffusion generation.

Notably, we observe that saturated images exhibit improved subject identity consistency. However, it is worth noting that this effect is primarily due to the saturation of colors, which conceals subtle variations in the subject's appearance across different views, thereby enhancing the perceived consistency. Furthermore, we observe that the quality of individual images remains high, as indicated by the FID metric, across our experiments. It's worth noting that our primary focus is elsewhere, and the FID metric may not fully capture the subtleties of consistency and variations in our results.

Sampling Saturation. To illustrate the effectiveness of our approach in removing the effect of saturation, we compared our results with those obtained using the original iterative sampling approach without identity token, as seen in Table 1. We found that our method consistently outperforms the original sampling method in terms of both quantitative metrics and visual quality of generated images. Specifically, the proposed method preserves the identity and consistency better while retaining the image quality.

Prompt Editing. As argued in the introduction, existing methods for 3D-aware novel view synthesis (StyleNeRF) lack the ability to guide the generation process based on prompts (or other conditioning factors). Existing diffusion models, do have this property, but they do not generate 3D-aware novel views (regular generation (Rombach et al., 2022; Ramesh et al., 2022; Midjourney.com, 2022), ControlNet (Zhang and Agrawala, 2023)) or they require a separate model for each instance (DreamBooth (Ruiz et al., 2023), DreamFusion (Poole et al., 2022), etc.). To the best of our knowledge, we present the first model which performs 3D-aware image synthesis, with inference-time text editing capabilities.

Our method allows for the generation of novel views and editing of real faces using a single input image. To achieve this, we estimate the camera position, angle, and subject identity using a pretrained projector. We then sample the NeRF to obtain the shape structure of the face, while the texture is estimated using the adapter *A* on regular diffusion sampling. These results can be seen in Figure 4.

The method enables quick editing of a generated or inverted real image while simultaneously generating new views. Because the learned conditioning plays a significant role in determining the image output, we make a slight adjustment in the frozen diffusion model. Specifically, we increase the weight of the desired new prompt tokens by approximately 1.5 times, aiming to enhance their influence on the generation process. Despite this modification, the output remains consistent, as shown in Figure 4 and 5.

5 CONCLUSIONS

We have presented a novel approach to generate new viewpoints from a single image by breaking down the complex task into sub-tasks and utilizing off-the-shelf pretrained architectures for each sub-task. Our approach enabled us to achieve high-quality results by leveraging the rich representation of these pretrained models and requiring only modest additional training. Specifically, we employed a pretrained NeRF model to estimate a robust shape prior, which we used to condition the pretrained diffusion model for image generation. We also introduced a consistency token that automatically maintained the desired details during the generation of new views, resulting in a cohesive relationship between style and shape. We identify and address an issue with style saturation during iterative sampling with classifier-free guidance. Finally, our proposed framework enables additional capabilities such as image-to-3D inversion, zero-shot image translation, and editing.

Limitations and Future Work. The main dependency of the method is the 3D-aware models used as shape prior, in terms of quality of the shape generated, as it affect most part of the consistency. Another limitation is the image generation quality of these 3D-aware models, since we do not use real data we are bounded to it when the consistency tightens between the input and output. Future work includes investigation of better priors, improving the generation quality beyond NeRF quality and better prompt edition.

ACKNOWLEDGEMENTS

This work is funded by the Spanish Government grants PID2022-143257NBI00, TED2021-132513B-I00, PID2019-104174GB-I00, MCIN/AEI/10.13039/501100011033 and by CERCA Programme / Generalitat de Catalunya.

REFERENCES

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. (2023). Universal guidance for diffusion models. *arXiv preprint arXiv:2302.07121*.

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L. J., Tremblay, J., Khamis, S., et al. (2022). Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion.

Gu, J., Liu, L., Wang, P., and Theobalt, C. (2022). Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*.

Han, I., Yang, S., Kwon, T., and Ye, J. C. (2023). Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*.

Haque, A., Tancik, M., Efros, A., Holynski, A., and Kanazawa, A. (2023). Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ho, J. and Salimans, T. (2021). Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hu, M., Zheng, J., Liu, D., Zheng, C., Wang, C., Tao, D., and Cham, T.-J. (2023). Cocktail: Mixing multi-modality controls for text-conditional image generation. *arXiv preprint arXiv:2306.00964*.

Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., and Zhou, J. (2023). Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Kamata, H., Sakuma, Y., Hayakawa, A., Ishii, M., and Narihira, T. (2023). Instruct 3d-to-3d: Text instruction guided 3d-to-3d conversion. *arXiv preprint arXiv:2303.15780*.

Karras, T., Laine, S., and Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *CVPR*.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023a). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. (2023b). Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*.

Liu, R., Wu, R., Hoorick, B. V., Tokmakov, P., Zakharov, S., and Vondrick, C. (2023). Zero-1-to-3: Zero-shot one image to 3d object.

Midjourney.com (2022). Midjourney. https://www.midjourney.com.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. (2023). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J. J., and Kemelmacher-Shlizerman, I. (2022). Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Schwarz, K., Liao, Y., Niemeyer, M., and Geiger, A. (2020). Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166.

Seo, J., Jang, W., Kwak, M.-S., Ko, J., Kim, H., Kim, J., Kim, J.-H., Lee, J., and Kim, S. (2023). Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*.

Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE.

Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., and Chen, D. (2023). Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior.

Wang, C., Chai, M., He, M., Chen, D., and Liao, J. (2022). Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3835–3844.

Wang, C., Jiang, R., Chai, M., He, M., Chen, D., and Liao, J. (2023). Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–15.

Wang, W., Yang, S., Xu, J., and Liu, J. (2020). Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing*, 29:9125–9139.

Xia, W. and Xue, J.-H. (2022). A survey on 3d-aware image synthesis. *arXiv preprint arXiv:2210.14267*.

Zhang, K., Riegler, G., Snavely, N., and Koltun, V. (2020). Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*.

Zhang, L. and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhou, P., Xie, L., Ni, B., and Tian, Q. (2021). Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*.