


Coloring 3D Avatars with Single-Image

Pin-Yuan Yang¹^a, Yu-Shan Deng², Chieh-Shan Lin³, An-Chun Luo² and Shih-Chieh Chang^{1,2,3}

¹College of Semiconductor Research, National Tsing Hua University, Hsinchu, Taiwan

²Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan

³Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan

Keywords: 3D Avatar, VR/AR Application, Deep Learning, 3D Model Generation.

Abstract: 3D avatars are important for various virtual reality (VR) and augmented reality (AR) applications. High-fidelity 3D avatars from real people enhance the realism and interactivity of virtual experience. Creating these avatars accurately and efficiently is a challenging problem. A lifelike 3D human model requires precise color representation. An accurate representation of the color is essential to capture the details of human skin, hair, and clothing to match the real people. Traditional methods, such as 3D scanning and multi-image modeling, are costly and complex, limiting their accessibility to an average user. To address this issue, we introduce a novel approach that requires just a single frontal image to generate 3D avatars. Our method tackles critical challenges in the field of single-image 3D avatar generation: color prediction. To achieve better prediction results, we propose a hybrid coloring technique that combines model-based and projection-based methods. This approach enhances 3D avatars' fidelity and ensures realistic appearances from all viewpoints. Our advancements have achieved better results in quantitative evaluation and rendering results compared to the previous state-of-the-art method. The entire avatar-generating process is also seven times faster than the NeRF-based method. Our research provides an easily accessible but robust method for reconstructing interactive 3D avatars.

1 INTRODUCTION

3D avatars have become essential for virtual reality (VR) and augmented reality (AR) applications, enhancing the realism in virtual communication, immersive gaming, and digital entertainment. Creating life-like avatars requires accurate 3D geometry and color representation, particularly for details such as human skin, hair, and clothing. Traditional approaches such as 3D scanning (Yu et al., 2021) and multi-image modeling (Jiang et al., 2022; Weng et al., 2022) produce high-quality results but are costly, complex, and time consuming, limiting accessibility to average users.

To simplify the process, single image-based methods (AlBahar et al., 2023; Cao et al., 2022; Feng et al., 2022; Huang et al., 2024; Qian et al., 2024; Saito et al., 2019; Saito et al., 2020; Xiu et al., 2022; Xiu et al., 2023) have emerged. These methods enable 3D avatar generation using a single frontal image. These methods eliminate the need for multiple images or

specialized hardware, making them efficient and user-friendly. However, a key challenge in single-image 3D avatar generation lies in coloring, particularly for unobserved viewpoints like the backside.

Single-image coloring techniques generally fall into two categories: model-based methods and projection-based methods. In this paper, we propose a hybrid coloring approach that combines model-based prediction with projection-based methods to address their own limitations. Our method retains fidelity while inferring unseen regions by predicting the backside image from the frontal input using a modified SPADE network (Park et al., 2019). Additionally, we introduce a region-specific loss function (RS Loss) to refine critical regions such as clothing, improving backside image quality. Our approach transforms the challenging 3D coloring task into a more manageable 2D problem, enabling efficient training using text-to-image data augmentation. Our coloring results achieve better quantitative results than state-of-the-art (SOTA) and a 7x faster generation speed than the multi-image modeling (NeRF-based) method.


^a <https://orcid.org/0009-0004-8528-4598>



Figure 1: Two example results from our 3D avatar generating system. The images from left to right are the input, front, and back views of the predicted avatar model. Input images were generated by DALL-E (Ramesh et al., 2021).

2 RELATED WORKS

2.1 Traditional 3D Reconstruction

Traditional methods, including 3D scanners and multi-image modeling, offer high-quality results. Scanners, such as those employed in (Yu et al., 2021), utilize 360-degree imaging to capture detailed 3D geometry but are expensive and time-consuming. Multi-image methods offer lower equipment costs than traditional 3D scanners, but still require significant time and computational resources for modeling (Weng et al., 2022). Many of these methods leverage neural radiation fields (NeRFs) to reconstruct 3D objects from multiple 2D images. NeRF-based approaches use neural networks to represent scene radiance and depth, allowing for detailed and realistic 3D reconstructions. However, these methods are not suitable for casual users due to their complexity and cost.

To address these issues, recent single-image-based approaches have sought to democratize 3D avatar creation by reducing the need for specialized equipment (Cao et al., 2022; Feng et al., 2022; Huang et al., 2024; Qian et al., 2024; Saito et al., 2019; Saito et al., 2020; Xiu et al., 2022; Xiu et al., 2023). Each single-image approach offers unique advantages. However, only a subset of works on single-image 3D avatar generation has addressed the coloring aspect of 3D avatars. Generating high-fidelity colors of a 3D avatar from single images can be particularly challenging compared to multi-image approaches due to the limited information available to generation models.

2.2 Single-Image Coloring

Single-image 3D avatar coloring methods fall broadly into two categories:

1. **Model-Based Methods:** Model-based techniques predict vertex colors or parametric models to reconstruct unseen regions (Cao et al., 2022; Saito et al., 2019; Huang et al., 2024; Qian et al., 2024). These methods utilize deep learning to infer plau-

sible 3D geometry and coloring from limited input data. Despite their ability to generalize well, they often require extensive datasets for training and struggle with high-fidelity backside generation due to limited information.

2. **Projection-based methods:** (Saito et al., 2020) project input images directly onto a 3D model in a pixel-aligned manner. This approach preserves the resolution and texture quality of the input image, but introduces artifacts, such as color duplication, particularly at the model’s edges and side areas.

Given these observations and challenges, our aim is to address their limitations while leveraging their strengths using novel hybrid approaches.

3 METHODOLOGY

This work aims to construct high-fidelity colored 3D avatars with a single-image model. To this end, we begin by synthesizing a 3D mesh model using PI-FuHD (Saito et al., 2020), a well-known model to construct 3D avatars from single images. This is followed by a novel coloring method, thereby achieving our goal of reconstructing a high-fidelity 3D avatar. The workflow of our avatar generation system is shown in Figure 2. Our coloring method directly projects the desired render result onto the 3D avatars. We need to acquire the backside view from the single-frontal image input. Once we have the backside image, we can integrate it with the frontal image to create a comprehensive color for the 3D avatars.

To complete this workflow, we have taken several key steps and innovations that collectively form our proposed method. First, we modified a well-designed image-to-image model, SPADE (Park et al., 2019), to generate the backside-color image of the target person based on the frontal view image. We introduce a region-specific loss function (RS loss) to refine the generation of the avatar’s backside (Section 3.1). After obtaining a predicted backside image, we project the front image and the predicted back image onto their respective sides of the 3D avatar in a pixel-aligned manner (Section 3.2). This direct projection method ensures that the color closely aligns with the contours and features of the 3D avatar mesh, enhancing its realism and fidelity.

Moreover, our strategy effectively transforms the 3D-generating challenge into a more manageable 2D task. This transformation enables the augmentation of training data through advanced text-to-image tools (Section 3.3). Moreover, our approach focuses on

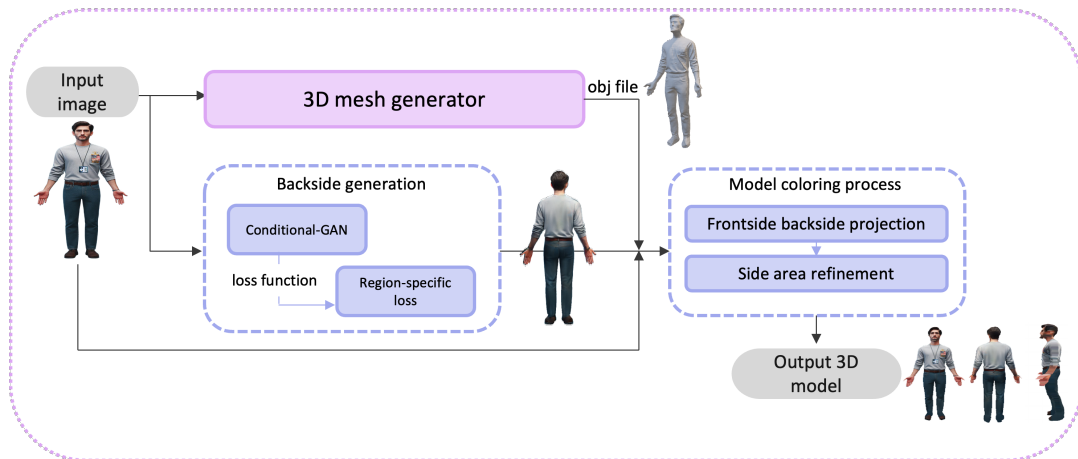


Figure 2: Proposed workflow for coloring single-image 3D avatars. We generate a non-colored avatar using PIFuHD (Saito et al., 2020) and predict the backside image with a generative model based on (Park et al., 2019). The complete avatar is formed by projecting both front and backside colors onto the 3D model.

generating avatars in a specific pose during our experiments, which aids further applications in whole-body skeleton rigging automation.

3.1 Backside Image Generation

In some previous works (Cao et al., 2022)(Huang et al., 2024)(Qian et al., 2024)(Saito et al., 2019), neural network models were utilized to predict the model’s color. Meanwhile, (Saito et al., 2020) projected the front image onto the model, resulting in duplication of the frontal content. To address this issue, we used an image-to-image model to predict the backside color image. We then projected the frontal and predicted backside image onto the 3D model. We selected SPADE (Park et al., 2019) as our base model due to its proven effectiveness and versatility. Although SPADE traditionally operates on semantic image synthesis, we have innovatively adapted it to process RGB image synthesis. The label-to-image capability of SPADE demonstrates its advantages for our task. Inherently, it understands and preserves the semantic features between images. For example, it recognizes similarities in colors and textures between the clothing on the front and back of the subjects. For instance, if the subject wears a blue shirt on the front, SPADE ensures that the generated backside image maintains a consistent color and texture for the shirt. This semantic consistency extends to other clothing items such as pants, ensuring a coherent and realistic appearance throughout the avatar. Our model architecture is shown in Figure 3.

However, during our experimental process, we observed some intriguing phenomena. Cloth regions often struggle with predictability because of differences

between front and back. For example, when a subject wore a jacket, the cloth underneath the jacket affected the generation of the backside image, as shown in Figure 4. This resulted in the backside appearing more similar to the front rather than reflecting the typical colors and textures one would expect on the back of the jacket. This circumstance prompted us to delve deeper into the model to handle such scenarios more

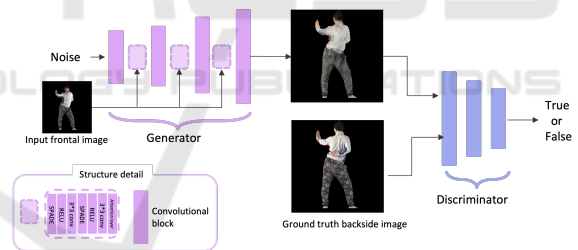


Figure 3: Our image-to-image model architecture. The SPADE blocks mention in structure detail are from (Park et al., 2019).



Figure 4: A visual example illustrates how the cloth underneath a jacket influences the appearance of the generated backside image. The input image were generated by DALL-E (Ramesh et al., 2021).

effectively.

The following formula is the objective function of the Pix2Pix generator (Isola et al., 2017), a GAN-based model.

$$\begin{aligned} L_{cGAN}(G, D) \\ = E_{x,y}[\log D(y)] + E_{x,z}[\log(1 - D(G(x,z)))] \end{aligned} \quad (1)$$

In this formula, G and D represent the generator(G) and discriminator(D) respectively. z is the noise factor we input when generating images. x is the frontal input image in our task. y is the output backside image we get from the G and y' is the ground truth. Based on this formula, the objective of G is to minimize the loss function. This means G aims to produce images that can mislead D into believing that they are real images. The G wants the D to be unable to distinguish between the generated images and the real images. Therefore, the generator strives to minimize the second term $E_{x,z}[\log(1 - D(G(x,z)))]$ (Discriminator loss). So, $D(G(x,z))$ is as close to 1 as possible, indicating that the generated images appear as realistic as the real images. Moreover, to further enhance the conditioning constraints, an additional mean squared error(MSE) loss $E_{x,y',z}[\|y' - G(x,z)\|^2]$ (GAN loss) is to specifically compare the generated image with the ground truth. This leads the generated images to resemble the ground truth as closely as possible.

To address the unique challenge we encountered in this task, we implemented a refinement in our approach. We introduced a region-specific constraint to the image-to-image model, focusing on the cloth area. The region-specific constraint is achieved by modifying the traditional loss function. For each input image, we utilize a human parsing extractor proposed by (Li et al., 2020) to segment the body of the target person, as shown in Figure 5. We then extract the cloth part from the results of the segmentation and turn it into a mask, denoted γ . The mask γ then serves as a guide to extract the exact part of the cloth region from the predicted (Mp) and the ground truth image (Mg).

$$\begin{cases} Mp = \text{Predicted image} \cdot \gamma \\ Mg = \text{Ground Truth} \cdot \gamma \end{cases} \quad (2)$$

The proposed RS loss is to measure the MSE between Mp and Mg . The new loss function comprises the existing generator, the discriminator loss, and the region-specific loss. The loss function is shown in Equation 3.

$$\begin{aligned} RS \text{ loss} \\ = MSE(Mp, Mg) + GAN \text{ loss} + Discriminator \text{ loss} \end{aligned} \quad (3)$$

We took corrective action since we observed a high incidence of errors in the cloth area. We added an extra loss function for this area to guide the model's attention and prioritize accurate predictions. As a result, this enhancement improves the model's ability to generate realistic backside images with accurate clothing details.



Figure 5: A visual example of our region-specific refinement. The right image is the segmentation result of the input image on the left. The purple area in the segmentation result represents our focused cloth area.

3.2 Projection Processing

In PIFuHD (Saito et al., 2020), directly projecting the image onto the 3D model leads to two main problems. First, determining which color map a vertex belongs to, either frontal or backside, poses a challenge. Second, side areas may experience mismatches that lead to discontinuities in the color of an avatar. To ensure the coherence and realism of 3D avatars, we adopt a projection strategy that leverages information from both frontal and back images. For each vertex (V), the normal vector (N) is computed relative to the camera vector (C). Here, we utilize the inner product of N and C to determine to which side V belongs. The inner product formula is shown as follows:

$$\vec{N} \cdot \vec{C} = |\vec{N}| \cdot |\vec{C}| \cdot \cos\theta \quad (4)$$

θ is the angle between N and C . According to the given formula, when θ is acute, $\cos\theta$ is greater than zero. We can conclude that the inner product will be greater than zero. Similarly, when θ is obtuse, $\cos\theta$ is less than zero. We can infer that the inner product will be less than zero in this case. More specifically, if the angle θ between N and C is acute ($\cos\theta \geq 0$), the vertex V is identified as belonging to the front area of the model. In contrast, if $\cos\theta < 0$, V is classified as part of the back area. With these computed inner

products, the vertices are categorized based on their orientation relative to the camera direction.

$$\begin{cases} \vec{N} \cdot \vec{C} \geq 0, & \text{the vertex is front} \\ \vec{N} \cdot \vec{C} < 0, & \text{the vertex is back} \end{cases} \quad (5)$$

Based on the vertex categorization, color information is selected from the appropriate image view. For front-facing vertices, the color values are extracted from the front image. For back-facing vertices, the color values are sourced from the predicted backside image. Both front and back are projected in a pixel-aligned manner. In this way, the front and back images can correctly project their pixels onto 3D avatars. In addition, we expand the edges of the front and back images. The expansion process ensures that every vertex can find a corresponding color value from the expanded images. This enhances the completeness of our color projection method. By integrating these multiple strategies, our projection process excels in reproducing the visual quality of 2D photos onto 3D avatars.

3.3 Addressing Dataset Limitations

The image-to-image model proposed by (Park et al., 2019) efficiently handled the task of predicting the backside image. However, we faced a significant challenge due to the limited availability of large-scale datasets that feature frontside and backside images of subjects in matching poses. Although some approaches, such as (AlBahar et al., 2021)(AlBahar et al., 2023), did not require paired datasets, they relied on at least two additional models during inference. In addition, an additional optimization process was necessary to improve the quality of the back image. These approaches did not align with our objective of making the overall process more accessible and efficient. Given these considerations, we chose an approach that required a paired dataset but achieved one-step generation of the backside image. Recognizing the requirement for paired datasets as a potential bottleneck in our research, we created a dataset to address this issue. We used standard text-to-image AI tools such as DALL-E (Ramesh et al., 2021). These tools allowed us to generate a synthetic dataset that met our requirements.

The prompt for the AI tool was crafted to emphasize the requirements of 'full body' and 'front and back' images. This ensured that the generated dataset accurately represented the needed perspectives for training. Example prompt: "A full body real person standing wearing a jacket, including both frontal photo and back full body photo, white background, real person."

4 EXPERIMENTS

4.1 Dataset

We collect a set of datasets generated using AI techniques (Ramesh et al., 2021)(Rombach et al., 2021). This dataset consists of 6,499 pairs of humans. Each pair shows a person standing in a specific pose, with images from both the front and back views. Out of these, 5,999 pairs were used as the training set, while the remaining 500 pairs were for the test set.

4.2 Implementation Details

For our experiments, we utilized PyTorch as our deep learning framework due to its generalizability. The experiments were conducted on an NVIDIA RTX 2080 Ti GPU. We implemented the image-to-image model and optimized it using the Adam optimizer with a learning rate of 0.01. The training images were first cropped to the bounding box of the human and then rescaled to a resolution of 512×512 . The training process involved feeding the set of 5999 pairs through the model over approximately 500 epochs.

For evaluation, we used the test set of 500 pairs to assess the model's backside synthesis performance in terms of MSE and other relevant metrics.

4.3 Quantitative Evaluation

4.3.1 Evaluation Metrics

To evaluate the quality of our results, we used several widely used evaluation methods. These included the structural similarity index measure (SSIM), the peak signal-to-noise ratio (PSNR), and the learned perceptual image patch similarity (LPIPS). SSIM is a metric that measures the similarity between two images. Consider luminance, contrast, and structure, providing a comprehensive assessment of image quality (Wang et al., 2004). A higher SSIM value suggested that the quality of the predicted image is closer to the ground truth or the reference image. PSNR measured the reconstruction quality by comparing the maximum possible power of a signal and the power of the corrupted noise. Higher PSNR values suggested better image quality with less distortion. LPIPS is a metric that measures the perceptual similarity between two images. Using deep neural networks, it models human visual perception, focusing on how humans perceive image differences. A lower LPIPS value indicates that the predicted image is perceptually closer to the reference image.

4.3.2 Comparison

We implemented Pose-with-Style (PWS)(AlBahar et al., 2021), PIFu (Saito et al., 2019), and Magic123 (Qian et al., 2024) to compare with our backside results. Quantitative comparisons are shown in Table 1, while visual comparisons are shown in Figure 6. Most methods perform well in frontal rendering, but exhibit less stability when rendering from the back. Therefore, we focus our experiments on the results of the back-side rendering. This decision was made to avoid potential biases associated with frontal rendering data. It also allowed us to better investigate and compare the performance of various methods, specifically in backside rendering. Our method outperformed all the others with an SSIM of 0.91, PSNR of 32.96, and LPIPS of 0.18. These results demonstrated our method’s superior performance in generating back-side images. Our method surpassed others in terms of image similarity and quality. Furthermore, our method showed the highest SSIM and PSNR values, indicating higher image quality.

Table 1: Backside comparison.

Evaluation methods	SSIM	PSNR	LPIPS
PWS	0.85	13.24	0.28
PIFu	<u>0.88</u>	<u>14.81</u>	<u>0.23</u>
Magic123	0.86	13.76	0.26
Ours	0.91	16.48	0.21

4.4 Ablation Studies

4.4.1 Data Augmentation

Due to the change in the task from 3D to 2D, we were able to significantly increase the training dataset. Initially, our training dataset was rendered from (Yu et al., 2021). However, the results were not satisfactory because of the small size of the dataset. To address this issue, we employ text-to-image models to generate a more extensive training dataset, thereby greatly enhancing the quantity and diversity of our training data. Table 2 and Figure 7 compare the performance before and after this substantial increase in training data.

Table 2: Data augmentation.

Evaluation methods	SSIM	PSNR	LPIPS
w/o data augmentation	0.89	15.25	0.22
w/ data augmentation	0.91	16.48	0.21

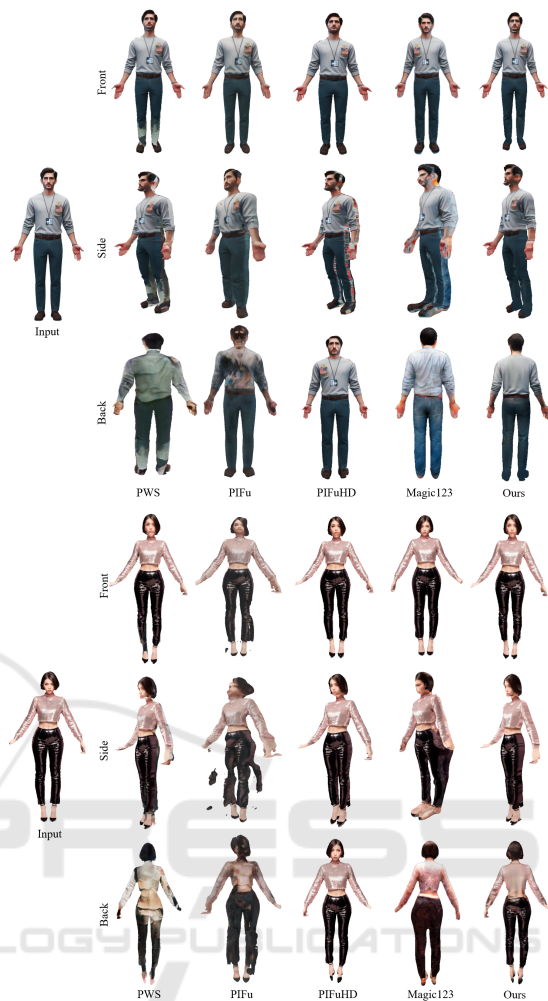


Figure 6: Comparison of our results with several previous works (AlBahar et al., 2021)(Qian et al., 2024)(Saito et al., 2019)(Saito et al., 2020). Our results demonstrate superior visual performance on both the side and back areas. Input images are from (Ramesh et al., 2021) and (Rombach et al., 2021).

4.4.2 Region Specific Refinement

To assess the impact of the region-specific refinement on our proposed method, we performed an ablation study comparing the performance with and without this refinement. The quantitative results are shown in Table 3, and the visual comparisons are in Figure 8. Without region-specific refinement, the model struggled to accurately predict cloth regions, leading to inconsistencies in cloth texture and appearance, especially in complex clothing items such as jackets. With region-specific refinement, the model performs better in generating realistic and coherent backside images, particularly in the cloth regions. The refinement helped the model focus on the cloth area and prioritize

accurate predictions, enhancing quality and realism.

Table 3: Region specific refinement.

Evaluation methods	SSIM	PSNR	LPIPS
w/o RS loss	0.91	16.21	0.21
w/ RS loss	0.91	16.48	0.21



Figure 7: Visual comparison of image generated with and without data augmentation. The images from left to right are, respectively, the input image, the result without data augmentation, and the result with data augmentation. The training dataset increased from 525 to 5999. The prediction quality significantly improved after data augmentation.



Figure 8: Visual comparison of image generated with and without RS refinement. The images from left to right are, respectively, the input image, the result without RS refinement, and the result with RS refinement. The quality significantly improved after applying RS refinement.

4.4.3 Side Area Refinement

We have compared the performance of the side area refinement before and after its implementation. The visual comparison is shown in Figure 9. We do not provide quantitative results in this ablation study since we lack the ground truth for the side area render results. However, based on the visual results, the side area artifacts are significantly reduced with the implementation of the refinement. The refinement visibly minimized these issues, resulting in smoother transitions and more coherent colors across the model’s surface.



Figure 9: Two visual comparisons of side area refinement before (left) and after (right) implementation. The images, after refinement, show a noticeable reduction in artifacts compared to their counterparts on the left. The deformities observed in the avatars are due to the imperfect generation by PIFuHD.

4.5 Limitation and Future Work

While our approach leveraged the reconstruction results from PIFuHD (Saito et al., 2020), it inherited any limitations or imperfections present in their reconstructions. This dependency on reconstruction quality could sometimes affect the quality of our coloring approach.

We aim to refine and expand our coloring method for future work to create more detailed 3D avatars. By doing so, we hope to generate 3D avatars with realistic colors that resemble real humans in shape and finer details. This will involve integrating high-fidelity reconstruction methods with advanced coloring techniques to achieve more accurate and lifelike 3D avatars.

5 CONCLUSION

We proposed a system that generates high-fidelity 3D avatars from a single image. This system addresses the main challenge in the field of single-image-generating models: color prediction. Our coloring method hybridizes the advantages of both previous model-based and projection-based methods to predict realistic color representations. We introduced a semantic-based region-specific loss function to enhance color prediction accuracy. In conclusion, our method outperformed the previous state-of-the-art approach by achieving better quantitative results and a 7x faster inference time compared to the NeRF-based method (Qian et al., 2024). Our system significantly lowers the barrier for the average user to reconstruct a 3D avatar in virtual space compared to the previous. Our research will contribute to vigorous development in various 3D avatar applications, facilitating greater accessibility and immersion in VR/AR

environments. This advancement improves the user experience and opens new possibilities for using 3D avatars in games, social networks, professional training, and virtual meetings.

ACKNOWLEDGMENT

This research was supported by the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 113-2640-E-194-001.

REFERENCES

- AlBahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., and Huang, J.-B. (2021). Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics*.
- AlBahar, B., Saito, S., Tseng, H.-Y., Kim, C., Kopf, J., and Huang, J.-B. (2023). Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia*.
- Cao, Y., Chen, G., Han, K., Yang, W., and Wong, K.-Y. K. (2022). Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2729–2739.
- Feng, Q., Liu, Y., Lai, Y.-K., Yang, J., and Li, K. (2022). Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In *NeurIPS*.
- Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., and Thies, J. (2024). TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.
- Jiang, W., Yi, K. M., Samei, G., Tuzel, O., and Ranjan, A. (2022). Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*.
- Li, P., Xu, Y., Wei, Y., and Yang, Y. (2020). Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.-Y., Skorokhodov, I., Wonka, P., Tulyakov, S., and Ghanem, B. (2024). Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.
- Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Saito, S., Simon, T., Saragih, J., and Joo, H. (2020). Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.
- Weng, C.-Y., Curless, B., Srinivasan, P. P., Barron, J. T., and Kemelmacher-Shlizerman, I. (2022). HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220.
- Xiu, Y., Yang, J., Cao, X., Tzionas, D., and Black, M. J. (2023). ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiu, Y., Yang, J., Tzionas, D., and Black, M. J. (2022). ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306.
- Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., and Liu, Y. (2021). Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.