

001

A. Implementation Details

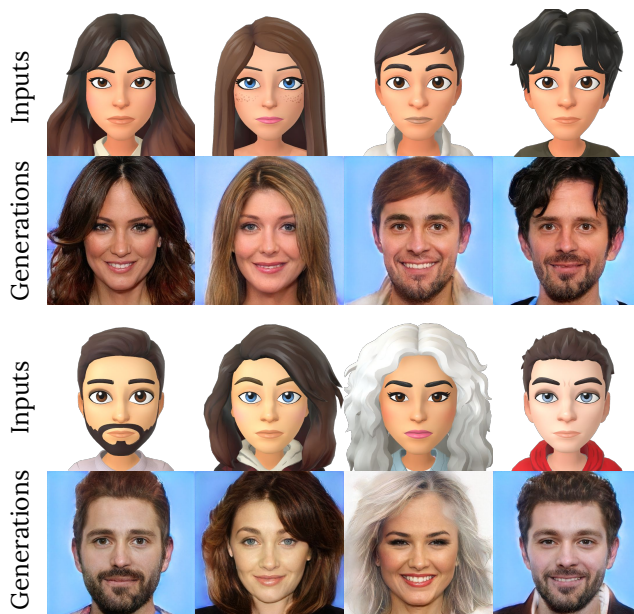


Figure 1. **GDA Training Data.** Visualization of data pairs used to train the GDA network. Each avatar is inverted into the latent space of a GAN, and a generator trained on realistic faces creates a corresponding realistic face, preserving features such as hairstyles, hair colors, and general facial characteristics.

002

A.1. Bitmoji Training Data

003

004

005

006

007

008

009

010

011

012

013

014

015

GDA Training Data. To train our Gaussian Domain Adaptation network, we start with a dataset of random Bitmojis and use GAN inversion to generate corresponding realistic images. Fig. 1 showcases some examples of the training data generated through this process. The resulting faces mirror the hairstyles, hair colors, and facial features of the original avatars. While the generated images may contain artifacts and exhibit limited diversity, our GDA model benefits from pre-training on Objaverse [1], enabling it to leverage prior knowledge and produce more detailed reconstructions than GAN inversion alone. This approach enhances the accuracy and expressiveness of the domain adaptation process.

016

017

018

019

020

021

022

023

Multi-view Training Data. Fig. 2 visualizes training samples from the Bitmoji dataset used for training our 3D Generation Network. Each avatar is rendered from 10 spherically distributed viewpoints around the head and is posed with random blendshape weights to simulate diverse facial expressions. The dataset features a wide range of hairstyles, skin tones, and accessories such as glasses, hats, and earrings. Although the U-Net is trained exclusively



Figure 2. **Multi-view Bitmoji Training Data.** Samples from the Bitmoji dataset used in training the 3D Generation Network. Avatars are rendered from the front and multiple random angles around the head, with random blendshapes applied to simulate various expressions.

on Bitmoji-style avatars, it effectively reconstructs dual-stylized avatars that exhibit distinct appearances and textures, demonstrating the network’s versatility and generalization capability.

024

025

026

027

A.2. Facial Action Coding System

028

We implement the following 16 blendshapes from the Facial Action Coding System. These blendshapes are compatible with most facial blendshape predictors like *Apple ARKit* or *Google Mediapipe*.

029

030

031

032

- browDownLeft
- browDownRight
- browUpLeft
- browUpRight
- eyeBlinkLeft
- eyeBlinkRight
- jawOpen
- jawLeft
- jawRight
- lipsPucker
- mouthFrownLeft
- mouthFrownRight
- mouthSmileLeft
- MouthSmileRight
- mouthStretchLeft
- mouthStretchRight

033

034

035

036

037

038

039

040

041

042

043

044

045

046

047

048

To animate the Bitmoji avatars for training data, we used the publicly available Bitmoji rig available here: <https>.

At inference time, we can use a real-time blendshape predictor like *ARKit* or *Mediapipe* to puppet the avatars from a real video. Please see the attached HTML gallery for a demonstration.

A.3. User Interfaces

To showcase the intuitive features of the Instamoji system, we present videos of the interface interactions available in the HTML gallery. The avatar generation interface, crafted with *Gradio*, enables users to effortlessly create dual-stylized avatars from their own photos. In addition, the blendshape editor, developed using *Viser*, allows users to pose their avatars in 3D by adjusting blendshape weights, thereby controlling facial expressions.

During the diffusion stylization process, we provide users with key parameters to balance identity preservation with style diversity. These controls are listed by their significance: 1) Style Transition Strength; 2) Edge Preservation Level 3) Identity Consistency Factor.

Style Transition Strength: This parameter, inspired by methods similar to SDEdit, regulates the extent of the stylization transition. Lower values enable the dual-stylized avatar to retain more details from the original single-styled avatar input.

Edge Preservation Level: This setting influences how accurately the system maintains the structure of the avatar by preserving the edges from the single-styled input.

Identity Consistency Factor: This controls the strength of identity features from the initial input photo, ensuring that essential facial characteristics remain recognizable.

We encourage users to view the videos in the HTML gallery to observe how these parameters affect avatar generation, enhancing both creativity and user experience.

B. Additional Results

B.1. Results Gallery

We invite you to explore the HTML gallery, which features videos of Instamojiavatars animated in 3D. Access the gallery by opening the `index.html` file in your web browser. The gallery includes the following highlights:

1. Dual-stylized avatars with dynamic facial animations displayed from various novel viewpoints.
2. A demonstration of the avatars' capabilities in facial puppeting for augmented reality applications.
3. Screen captures of the Instamojiuser interfaces, showcasing the ease of creating dual-stylized avatars and posing them using blendshapes.



Figure 3. **Augmented Reality Puppeting.** This example demonstrates the use of *Mediapipe*'s real-time face detection to animate avatars based on estimated blendshape weights. By alpha-compositing the avatars with the original input, we enable dynamic puppeting in augmented reality. For live demonstrations, please refer to the HTML gallery.

B.2. More Applications

3D Avatar Animation. Dual-stylization offers the ability to swiftly visualize avatars in various scenarios, unlocking numerous applications. As illustrated in Fig. ??, Instamoji avatars can be employed to create personalized comics and stickers, offering users a unique way to express themselves. Another promising application lies in augmented reality (AR), where avatars can be controlled and animated with real-time tracked facial expressions. Examples of this application are shown in Fig. 3 and within the HTML gallery. By utilizing *Mediapipe*'s real-time blendshape tracker, we animate the 3D avatars and seamlessly integrate them with video content, enabling them to be rendered in an AR environment through alpha compositing.

Real-time Web Rendering. Our choice to represent avatars using Gaussian Splats enables efficient real-time rendering on mobile devices. As demonstrated in Fig. ?? and in the HTML gallery, the avatars achieve a rendering rate of 90-100 FPS on a laptop, and 30-40 FPS on a phone. When paired with a face tracker, these avatars can be used to generate engaging filters and augmented reality effects. The demonstration showcases an avatar rendered in Google Chrome on a MacBook, entirely on the client side.

GDA Generalization. GDA demonstrates that the features learned from few-shot 3D reconstruction models are transferable to new tasks. Shown in Fig. 4, GDA can be applied for more domains such as cats. We hope that GDA can in-



Figure 4. **GDA Generalization Across Domains.** This illustration showcases the versatility of Gaussian Domain Adaptation (GDA) as an image-to-image translation method. Demonstrated above is GDA’s capability to transform realistic cat images into anime-style representations and vice versa, highlighting its potential for a wide range of applications beyond avatar creation.



Figure 5. **Ablation Study on 3DMM Tracking.** This figure demonstrates the effects of using 3DMM features in conjunction with FACS blendshape weights. The combination enhances the expressiveness and fidelity of avatar animation, accommodating both realistic and stylized facial expressions.

spire future work on using Gaussian features for other tasks.

C. Ablation Studies

C.1. Gaussian Domain Adaptation

For the task of 2D image-to-image translation, we find that the LGMs, which generate 3D Gaussian splats, are still very

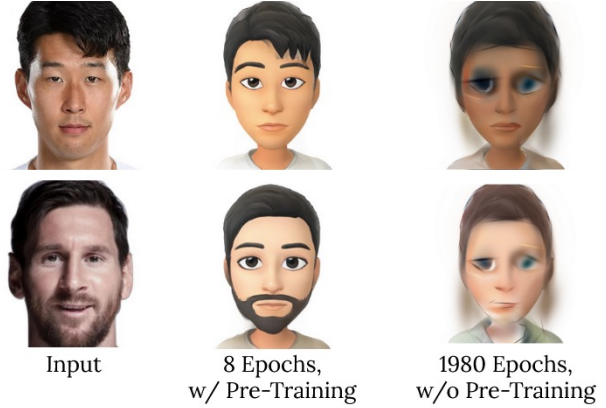


Figure 6. **Ablation Study on Pre-training.** Objaverse pre-training is necessary to obtain high-quality results from GDA. Without pre-training, the GDA model struggles to converge, even after 250x more epochs than with pre-training.

well-suited for this task. We believe this is due to two factors: that the LGM model is pre-trained on Objaverse, and that 3D-awareness helps create semantic correspondences in the 2D style transfer task. Shown in Figure 6, even after training a LGM model from scratch for 1980 epochs, it still does not converge. In addition, 3D-awareness helps generate accessories like sunglasses, which intuitively should be placed in front of a face. These two priors help the GDA model overcome the quality issues with the data in Fig. 1, leading to the high-quality results shown in the main text..

C.2. 3DMM Tracking

As shown in Fig. 5, our ablation study highlights the complementary strengths of 3DMM tracking and FACS-based blendshape features in avatar animation. 3DMM is adept at capturing realistic facial expressions, making it ideal for animating real faces, but it struggles with the exaggerated features typical of cartoon avatars. Conversely, FACS blendshapes excel in stylizing facial elements, such as eye and mouth shapes, crucial for cartoon animation. By integrating the precision of 3DMM with the expressive capability of FACS blendshapes, we enhance the overall animation quality, enabling our avatars to faithfully portray both realistic and stylized expressions, thus delivering a more versatile and convincing animation experience.

D. Ethical Discussion

The use of photorealistic avatars has raised significant privacy and ethical concerns, particularly in relation to their potential misuse in creating deep fakes and spreading misinformation. In contrast, stylized cartoon avatars offer a safer alternative as they are not easily exploited for direct impersonation. In our work, we have prioritized user privacy by ensuring that no real person’s images are used to

161 train our models. Instead, the realistic images employed
162 for training the Gaussian Domain Adaptation (GDA) sys-
163 tem are generated by a GAN. We recognize, however, that
164 GAN-generated data can reflect the biases present in the
165 original datasets used for training. Consequently, we re-
166 main vigilant about these limitations and are committed to
167 continuous evaluation and improvement to mitigate any un-
168 intended biases.

169 References

- 170 [1] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs,
171 Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani,
172 Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A uni-
173 verse of annotated 3d objects. *2023 IEEE/CVF Conference*
174 *on Computer Vision and Pattern Recognition (CVPR)*, pages
175 13142–13153, 2022. [1](#)