StyleMorph: Disentangled 3D-Aware Image Synthesis with a 3D Morphable StyleGAN - Supplemental

Eric-Tuan Le[1]* Edward Bartrum[1,2]* Iasonas Kokkinos[1]

[1] University College London [2] Alan Turing Institute

In this supplemental, we first demonstrate the quality of the dense correspondence learnt for each instance through the template (Section 1) by propagating semantic labels across instances. In Section 2, we show how our 3D-morphable model can be used to invert real images and deform the underlying shape while keeping the same appearance. Then, we evaluate the quality of our disentanglement both quantitatively and qualitatively (Section 3 and 4). We show state-of-the-art disentanglement between shape and appearance, and view consistency for given instances. In Section 5, we give more details on the alpha IoU consistency introduced in the paper.

In Sections 6 and 7, we provide additional ablations to show the benefit of TOCS compared to other conditioning signals (depth, rgb, neural features). Then, we demonstrate shape control over gender and facial expressions in Section 8. The equivariance of the TOCS is shown in the case of a simple rigid deformation in Section 9. Then, we present in Section 10 the limitations of our model and we introduce potential directions to mitigate them.

In Section 11, we show more qualitative results from our model. Then, we show additional results showing shape controllability for AFHQ categories (Cats, Dogs, Wild). We report the reader to the additional video at https://stylemorph.github.io/stylemorph/ to best view our synthesized 3d models in Section 12.

We finally provide additional details on our training details for reproducibility (Section 13). We will also publicly release the code.

1 Dense correspondence

Through our deformation network, we automatically learn dense correspondences between all instances generated by deforming our template (without explicit supervision). We evaluate the quality of the learnt correspondence in Fig. 1 where semantic labels of human faces are transferred from an image of one instance across all instances of the category. We use an off-the-shelf segmentation network (Yu et al., 2018) to label each pixel of an RGB image of the mean shape and propagate the label to all instances in Fig. 1. We show that the semantic segmentation stays very accurate even when large deformations are applied to the template (e.g growing hair). Our work could be used to generate synthetic datasets to improve semantic segmentation or keypoints prediction.

2 GAN Inversion

We show in Figure 2 how real images can be inverted using Pivotal Tuning (Roich et al., 2021) to recover the 3d model from a single image.

We process the inversion in two steps. First, our volume renderer is inverted to obtain the correct 3D shape from the input test image. To do so, we first predict the segmentation mask of the given image with the unsupervised *Labels4free* (Abdal et al., 2021) approach. Then, we optimise our volume renderer to find the correct pose (elevation and azimuth (ϕ, θ)) and shape code $\mathbf{z_s}$. Our DNR is then inverted using PTI (Roich et al., 2021) to obtain the remaining foreground $\mathbf{z_{fg}}$ and background code $\mathbf{z_{bg}}$ while locally updating the DNR weights.

Project page: https://stylemorph.github.io/stylemorph/

*Equal contribution.

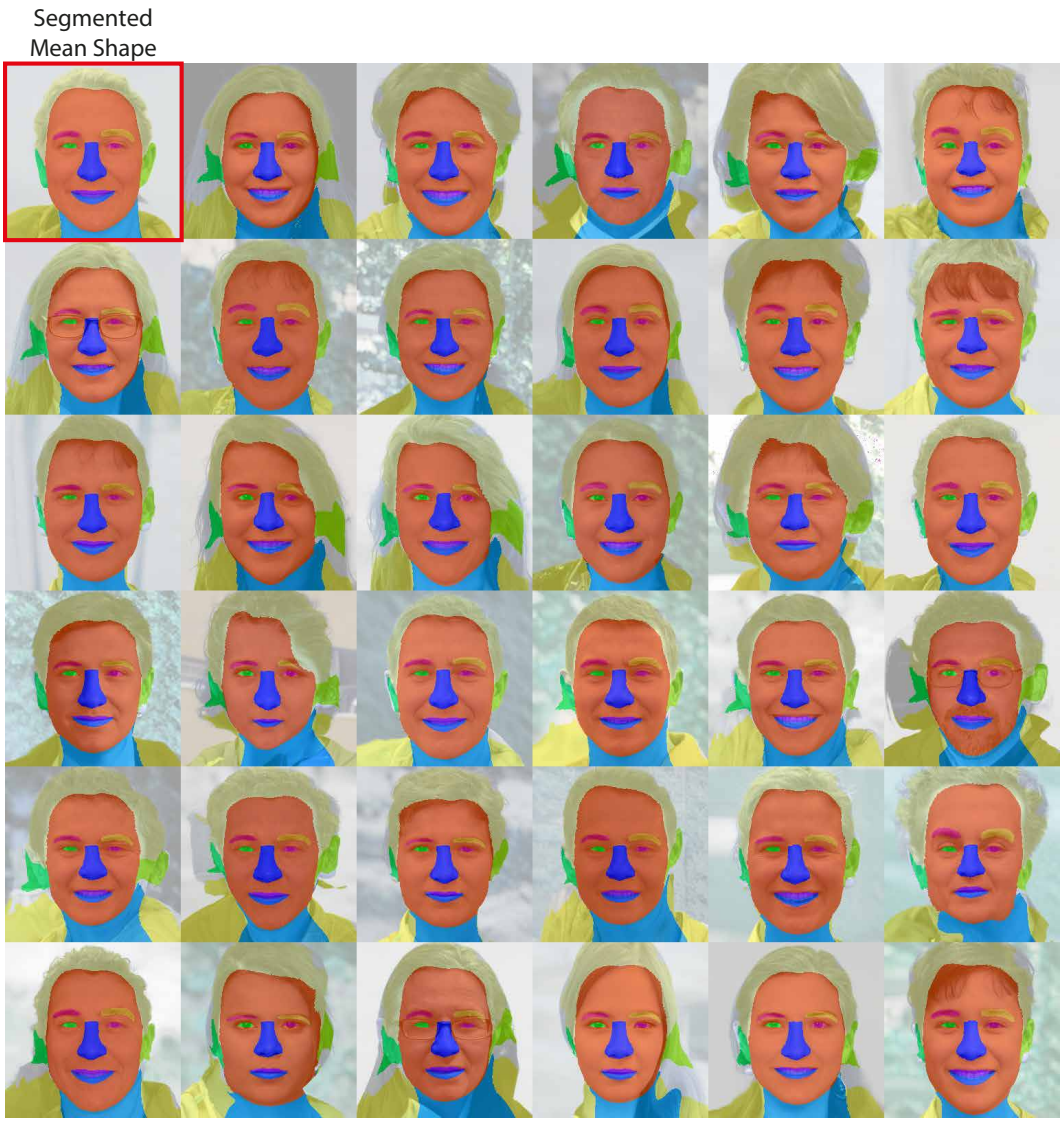

Figure 1: By learning to deform a template shape, our approach provides dense correspondence without explicit supervision. We demonstrate the dense correspondence by transferring the semantic label from one image (top left) to multiple other instances through the template.

Due to our unique deformation equivariant model, we can easily morph the original shape into a new random shape while keeping the initial appearance fixed.

3 Comparison with other controllable methods

Following Tewari et al. (2022), we evaluate the quality of our disentanglement in Table. 1 using their newly defined metrics. We compare consistency metrics and FID scores at multiple resolutions with 3 state-of-the-art controllable synthesis methods (Tewari et al., 2022; Xue et al., 2022; Chen et al., 2022). Note that SofGAN (Chen et al., 2022) synthesises images conditioned by semantic segmentation maps, but is controllable via a separately learned Semantic Occupancy Field.

We note that our Morphable Renderer differs from Tewari et al. (2022) as we are capable of working directly on a set of aligned images, without requiring any ground truth data. In order to ensure that our Morphable Model produces accurate deformations, we incorporated several shape losses to prevent degenerate cases. Additionally, we incorporated a 4-channel RGBA discriminator to guide the model towards generating realistic shapes via deformation. In contrast, Disentangled3D (Tewari

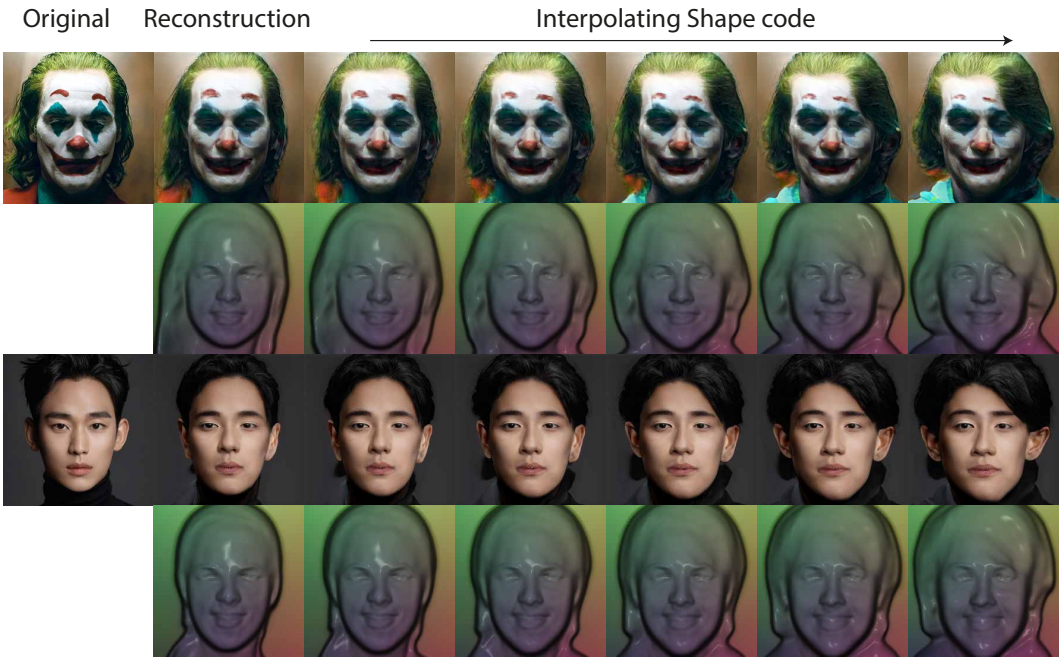

Figure 2: Real images of *The Joker* and *Kim Soo-hyun* are inverted using Pivotal Tuning (Roich et al., 2021) and then morphed into a new shape via offset interpolation.

	Geometry ↓	Appearance ↓	FID 256^2 ↓	FID 512^2 ↓	FID 1024^2 ↓
Disen3D (Tewari et al., 2022)	**0.39**	0.05	28.18	-	-
GiraffeHD (Xue et al., 2022)	0.62	0.09	11.93	-	10.13
SofGan (Chen et al., 2022)	0.82	0.03	-	12.79	-
Ours	0.81	**0.02**	**7.90**	**9.66**	**10.01**

Table 1: FFHQ Disentanglement Consistency and FID scores at 256^2, 512^2, 1024^2 resolutions. The appearance consistency measures how consistent is the appearance when only changing the shape (lower is better). The geometry consistency similarly measures how consistent is the geometry when changing the appearance (lower is better). In an apples-to-apples comparison, our method is best overall, providing photorealistic synthesis across all resolutions with state-of-art controllability.

et al., 2022) has only been tested on images with white backgrounds. Through the use of ground truth mask information, their model can be trained more simply without regularizing the shape model. However, the availability of ground truth segmentation mask is not guaranteed for more complex instances like in AFHQ. Since the Disentangled3D source code is not available, we were unable to test its performance on real FFHQ images with background for comparison.

First, the appearance consistency is measured as the standard deviation of the average color in a semantically well-defined region, when changing the shape while keeping appearance fixed. In our case, for FFHQ, we use the hair area obtained via an off-the-shelf semantic model (Yu et al., 2018). The lower the value, the more consistent is the appearance when changing the shape. We perform on par with SofGAN (Chen et al., 2022) while achieving stronger consistency compared to Disentangled3D (Tewari et al., 2022) and GiraffeHD (Xue et al., 2022).

Then, the geometry consistency is measured as the standard deviation of 66 facial landmarks obtained via an off-the-shelf tool (Saragih et al., 2011) when changing appearance while keeping the geometry fixed. The lower the number, the more consistent the shape is when varying the appearance.

Our FID scores are the strongest at all 3 resolutions. Our model performs best overall compared to the other controllable synthesis methods.

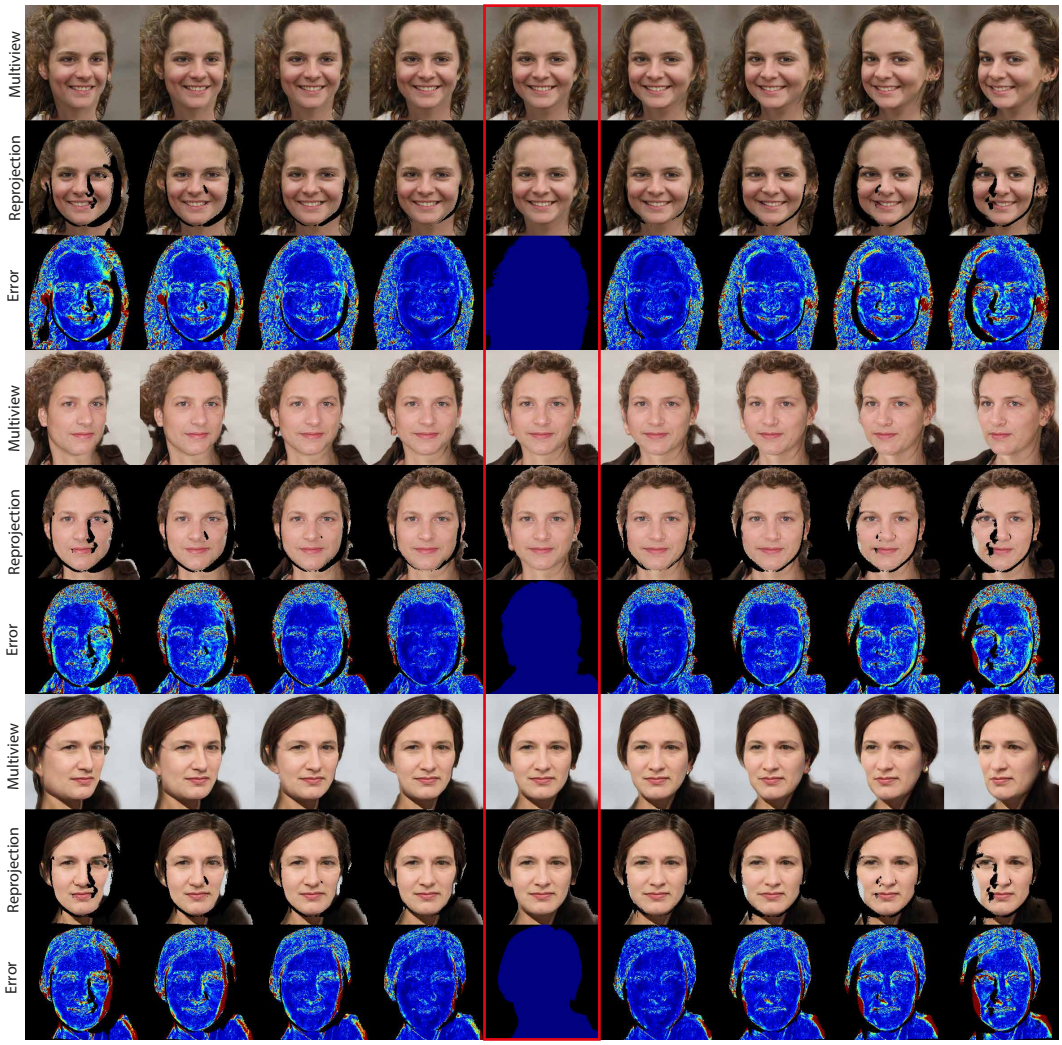

Figure 3: RGB rendering view-consistency. We render multiple views of the same instance and reproject the pixels to the frontal view (red column). We present the pixelwise error map to show the similarity between the frontal view and the side-view reprojections. The few reprojection errors are located in high frequency region of the images such as hairs, eyes and teeth.

4 View consistency

In Fig. 3, we demonstrate the quality of our pose disentanglement, showing strong view consistency. We present how side-view RGB images (1.5 x the standard azimuth deviation) reproject into face view. Our reprojected images closely match the frontal view demonstrating strong 3D consistency. As shown on the error maps, most of the errors are located in very high frequency areas such as the hair, eyes and teeth.

We quantify our model's view consistency by calculating the per pixel L2 error when reprojecting side views to frontal views. To make our metric robust to outliers, we report the median pixel RGB error across 1000 generated FFHQ object instances. We compare the view consistency scores for our various model variants in the ablation table (Table. 2) of the main paper.

5 Alpha IOU consistency

At test time, we evaluate the consistency of our model's generated image samples with their underlying alpha masks. For generated images, we use an off-the-shelf face segmentation system to

Full model	Depth map	Low-res RGB	RGB upsample (Wang et al., 2021b)
8.31	20.19	25.08	55.64

Table 2: FID scores ↓ on FFHQ 256^2. We compare our model's Deferred Neural Renderer and TOCS map conditioning signal with 3 baselines; 2 alternative conditioning signals, and a state-of the art rgb-upsampling GAN (Wang et al., 2021b). We find that our model's TOCS-based deferred neural rendering pipeline is critical for synthesising photorealistic images.

estimate their alpha masks (Yu et al., 2018). We calculate the IOU between the alpha mask sampled by the generator with the mask predicted by the segmentation network (higher is better). We report alpha mIOU scores for our model variants in the ablation table Tab 2. of the main paper.

6 Additional DNR ablation results

In order to further validate the merit of our TOCS conditioning signal, we perform additional ablation studies beyond the TOCS vs NOCS ablation originally performed in the main paper. We report FID results in Table. 2. We first replace our TOCS map input to the DNR with a simple camera-space depth map. We find that this geometric signal is less informative for the DNR since it does not provide surface-level correspondences, and is not consistent across viewpoints. Consequently, it results in lower quality image synthesis. We also try using the low-resolution RGB samples generated during stage-1 training by the volume renderer as a conditioning signal. Since the low-resolution RGB signal varies across different shape codes, this provides a less consistent conditioning signal which is again harmful to our model's FID score. Finally, we replace our Deferred Neural Renderer with an out-of-the-box image upsampling GAN (Wang et al., 2021b), and pass our low-resolution RGB samples to that. Again we find that the image quality is lower than what we report for our full model. All 3 ablations are evaluated on FFHQ 256^2.

7 Comparison with neural feature-based disentanglement

In Fig. 4, we show the effect of passing separate shape and appearance style codes to a neural-features based variant of our network. Instead of passing the TOCS to our DNR, we render neural features predicted by the SDF network, and pass them directly to the StyleGAN blocks of our DNR. We use a shape code to condition the implicit field which outputs the 2D neural features, and pass in a separate appearance code to the 2D StyleGAN blocks. We find that the resulting model exhibits entanglement of shape and appearance (top row); changing the shape code whilst keeping the appearance code fixed results in changes to both shape and appearance. We believe that this is due to the leakage of appearance information into the learned weights of the implicit field, which is not constrained to predict shape only. In contrast, our full model successfully separates shape and appearance, as the implicit field is only used to predict the TOCS map, which is purely geometric and cannot contain appearance information.

8 Shape-based gender and expression control

The motivation behind our work is the interpretability and editability offered by the classical 3D Morphable Model for human faces (Blanz & Vetter, 1999). A hallmark of that model has been the ability to modify the 3D shape of a synthesized face by editing the underlying shape code - allowing to modify the gender, expression and identity of a user.

The union of all such sources of variation represents the non-rigid variation of human faces in 3D (what we collectively refer to as "shape" in our work). Even though in the original work of Blanz & Vetter (1999) human annotations were used to identify directions related to gender and expression, in our work we only have unstructured sets of images, making it impossible to disentangle 3d deformable shape into its constituent factors. Still, we show interpolation results between different expressions by heuristically picking source and target latent codes that reflect a change in only one factor (expression/gender) as described below.

Figure 4: Top: Using a variable shape code to condition implicit field based neural features and a fixed appearance code for the 2D synthesis indicates entanglement. Bottom: Our full model is capable of disentangled shape synthesis.

We first demonstrate that our shape deformation model is capable of varying facial expressions using our FFHQ trained generator (Fig. 5). Although not trained using facial expression supervision, we can identify deformation code pairs in the model's latent deformation-space which correspond to changing expression. We select pairs with minimal changes to the head-shape in order to preserve identity. By interpolating between the shape codes with fixed appearance codes and pose, we can generate samples with smoothly varying facial expressions.

We similarly show qualitative expression deformation results for our model trained on the kaggle subset of the AffectNet (Mollahosseini et al., 2019) face image dataset in Fig. 6. This dataset consists of frontal face images featuring a greater range of facial expressions than those found in FFHQ.

We similarly demonstrate deformation-based control over the gender of generated samples from the our FFHQ trained model in Fig. 7.

9 NOCS VERSUS TOCS

In Fig. 8 we show what would be the output NOCS and TOCS if we were to apply rigid deformations in the object space. We show that the TOCS is equivariant to rigid deformation - each surface point keep the same TOCS value - while the NOCS is changed.

10 LIMITATIONS

10.1 RESULTS ON MAN-MADE OBJECTS

In order to test the limitations of our model we train it on two categories of man-made objects that have topological variability - which means that not all instances of the 3D category can be expressed in terms of a diffeomorphic warp from a 3D template to the instance. In such cases the "morphing" assumption breaks down and we observe that the model mitigates for this by absorbing structure variation into the appearance component. Even though this is geometrically "wrong", appearance-wise it can still look correct, and still enjoy the geometric controllability properties offered by the underlying morphable model.

In particular we show controllability results for our model on more general object categories (Cars and Buildings) in Fig. 10 and Fig. 11. We train our full model on the CompCars and Architecture datasets, and show disentanglement results across all 4 factors of control.

We observe for instance for cars that even though the 3D shape is coarse, it can deform from the outline of a convertible to a sedan - the other details are painted onto the car by the DNR.

For buildings the results are even more interesting - there is more substantial structure variability in these cases, with facades comprising different numbers of rows and columns, as well as different

Figure 5: Facial expression interpolation using our generator's shape deformation network. We select pairs of deformation-codes corresponding to matching identities, with different expressions. Then we interpolate between the deformation-offsets and generate samples whilst keeping the other latent codes fixed. Note the smiles widening from left to right, exposing the teeth, and the movement of the eyebrows in the top row.

numbers of buildings being present in the same photo. From the second, "shape" column we observe that the shape code controls the coarse outline of the buildings and their relationships, while from the "appearance" column we observe that appearance controls the number of floors and columns. The TOCS conditioning signal seems to have a vertical and horizontal periodicity - we conjecture that this acts like a scaffold, allowing the synthesis of periodic textures that deliver a consistent appearance across columns/floors.

We would not advocate using our model in its current form in order to learn the 3D shape of such categories. But we point out to some exciting recent works (Liu & Liu, 2020; Wang et al., 2022; Duggal & Pathak, 2022) that indicate one can learn morphable models with topological variability by expressing them through 3+D - dimensional implicit networks - with D controlling the degree of topological variability. We will consider using such models in future work.

Figure 6: Facial expression interpolation using our generator trained on the kaggle subset of the AffectNet (Mollahosseini et al., 2019) facial expression dataset. As with FFHQ (Fig. 5) we interpolate between deformations corresponding to varying expression, whilst keeping the appearance code fixed. Note the movement of the eyebrows and corners of the mouths, indicating changing expression.

11 Additional qualitative results

We show caricature results for Humans Fig. 9. We amplify the effect of deformation-field samples in order to generate shapes with exaggerated features. We observe that our DNR is robust to conditioning by these extreme deformations, even though they are not present within the sample space of the converged shape model.

We show additional qualitative shape control results on AFHQ Dogs (Fig. 12), Cats (Fig. 13), and Wild (Fig. 14) datasets. In each case, we keep the appearance code fixed whilst varying the shape code in order to show the diversity of shapes which can be generated by the model. We note that the appearance remains consistent whilst the shape changes substantially, e.g causing the ears to be stretched and folded, twisting and stretching the heads relative to the torso. We show the generated images alongside their underlying TOCS maps which they were conditioned by, to demonstrate the high degree of consistency between them.

We also show multiview consistency results on the challenging MetFaces dataset in Fig. 15. Since this is a small dataset, similarly to Zhou et al. (2021); Gu et al. (2021) we perform stage-2 DNR training on it using a stage-1 model which has been pretrained on the much larger FFHQ dataset.

Figure 7: Each row shows samples with fixed foreground and background appearance codes. We choose shape codes to synthesise male faces in the left group of 3, and female faces in the right group of 3. Note that our shape deformation network is capable of synthesising both male and female hairstyles and face shapes.

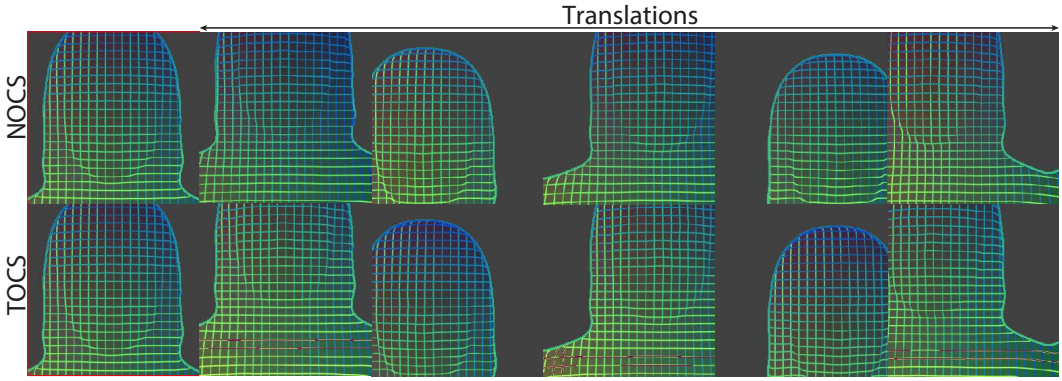

Figure 8: Effect on NOCS and TOCS contour maps, when applying a rigid deformation to the template shape. We apply multiple rigid translation deformations to the template shape (left column). We show the projected NOCS and TOCS with their contour maps. TOCS values (bottom row) shift equivariantly with the deformation, while the NOCS values (top row) are fixed in world space.

Finally, we also compare in Figure 16 the visual quality of our model compared to 3 state-of-the-art 3D-aware GANs. Our model is able to generate comparable photorealistic results while providing full control over multiple sources of variation.

12 VIDEO RESULTS

Our results are best viewed on video. On our project video, at https://stylemorph.github.io/stylemorph/ we show controllability results for pose, shape, foreground/background appearance. We show the effects of interpolating codes in the latent space for each factor of variation, on the FFHQ and AFHQ datasets. We also show the strong correspondence between TOCS maps and RGB synthesis results.

Figure 9: Deformation-based caricatures using our FFHQ-trained generator. We sample shape and appearance codes to generate the samples shown in the white column. Then, fixing the appearance code on each row, we show the result of interpolating deformations from the learned template (left column) up to the sampled shapes. In the rightmost columns, we interpolate the deformations further beyond the generated samples, synthesising shapes with exaggerated facial features. Note the rectangular head shape in the 2nd and 4th rows, and the enlarged hair on the other rows.

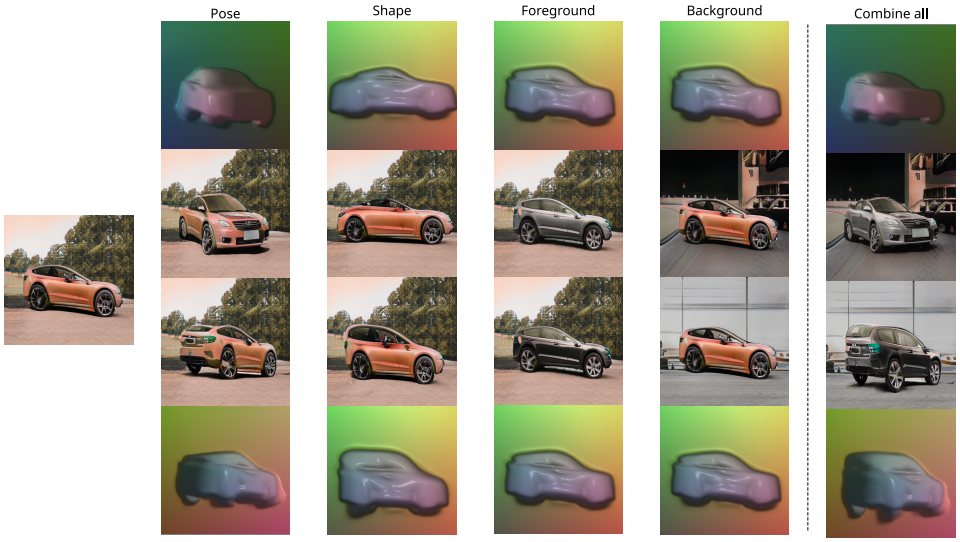

Figure 10: Disentangled image synthesis on the CompCars dataset. Starting from a synthesized sample on the left, we change one factor at a time, and on the right we show the compounded effect of changing them all.

13 Training details

13.1 Training the Morphable Volume Renderer

During the first stage of training, we bootstrap the Morphable Volume Renderer by training it to render RGB radiance fields as a low-resolution 3D aware generative image model. This is done in

Figure 11: Disentangled image synthesis on the Architecture dataset: starting from a synthesized sample on the left, we change one factor at a time, and on the right we show the compounded effect of changing them all.

Figure 12: Shape results on AFHQ Dogs at 512^2. Diverse shape deformations with a fixed appearance code. We observe that the deformation network can adjust the width of the dogs's head, enlarge its nose and fold its ears up or down. Results are best viewed on video.

order to learn a morphable-template shape model needed to produce the 2D TOCS maps which drive the 2D image synthesis network during the second stage of training.

This approach allows us to capture substantial amounts of shape variability through a generic SIREN-parametrized nonlinear function that generalizes the explicit linear parametric models of shape deformation (Kanazawa et al., 2018; Kulkarni et al., 2020; Kokkinos & Kokkinos, 2021), while being entirely trainable from raw 2D RGB images. In the volume-renderer stage of training, we render RGB values to generate low-resolution samples, whereas in the DNR stage, we instead render the template coordinates of the point where each ray intersects the object surface, which we call TOCS maps. The DNR can be trained at arbitrary high-resolutions conditioned on fixed-size TOCS maps (In our experiments, we always use TOCS maps of size 64^2).

13.1.1 Unsupervised approach to obtain real image segmentation masks

During the first stage of training, we rely on estimated alpha segmentation masks of the real image dataset to feed RGBA images to the low-resolution discriminator. To obtain the masks from raw

Figure 13: Diverse shape deformations with a fixed appearance code for AFHQ Cats at 512^2. We observe that the deformation network can twist the cat's head relative to the torso and adjust the position of its ears. Results are best viewed on video.

Figure 14: Diverse shape deformations with a fixed appearance code for AFHQ Wild at 512^2. We observe that the deformation network can twist the lions ears inwards or outwards and vary the distance between its eyes and the tapering of the nose. Results are best viewed on video.

RGB images, we start by training Labels4free (Abdal et al., 2021), a StyleGAN-based unsupervised method to generate synthetic images together with their predicted foreground-background segmentation mask. Then, to get the segmentation prediction for real images with Labels4free, we leverage the PSP (Richardson et al., 2021) encoder to project them into the StyleGan2 latent space. We then feed the predicted latents to Labels4free to get the segmentation mask for the corresponding real images.

13.1.2 Volume Rendering equations

In order to render the TOCS map (noted $\mathrm{TOCS}(\mathbf{r})$), we integrate the template-coordinate values $\hat{\mathbf{r}}(t)$ along each ray - each ray corresponding to a different pixel:

$$\mathrm{TOCS}(\mathbf{r}) = \int_{t_n}^{t_f} w(\hat{\mathbf{r}}(t))\hat{\mathbf{r}}(t)\mathrm{d}t \tag{1}$$

with $w(\hat{\mathbf{r}}(t))$ the SDF-based weight computed at the template point and t_n and t_f the near and far ray.

Figure 15: We show multiview consistency results on the challenging MetFaces dataset, with variable azimuth. We observe that our model demonstrates accurate 3D consistency even on artistic datasets which are not constrained to precisely represent 2D projections of 3D scenes.

To derive the SDF based weights - following Or-El et al. (2021) - we firstly compute the occupancy $\sigma(x)$ at each template point:

$$\sigma(\hat{\mathbf{r}}(t)) = \frac{1}{\alpha} \cdot \text{Sigmoid}\left(\frac{-d(\hat{\mathbf{r}}(t))}{\alpha}\right) \tag{2}$$

with $d(\hat{\mathbf{r}}(t))$ the predicted SDF value at template point $\hat{\mathbf{r}}(t)$ and α a learned parameter that controls the tightness of the density around the surface boundary.

Similarly to previous works, the SDF based weights can be simply computed from the occupancy:

$$w(\hat{\mathbf{r}}(t)) = T(t)\sigma(\hat{\mathbf{r}}(t)) \tag{3}$$

with $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\hat{\mathbf{r}}(s))ds\right)$.

We note that the integrals are approximated via discrete uniform sampling. We sample N uniform points between $[t_n, t_f]$:

$$t_i = \frac{t_f - t_n}{N} \cdot i + \delta \quad \text{with} \quad i \in \{0, ..., N-1\} \tag{4}$$

with δ a random offset sampled uniformly between $\left[0, \frac{t_f - t_n}{N}\right]$

13.1.3 Initialization of the Morphable Volume Renderer

Similarly to Or-El et al. (2021), we initialize the volume renderer to predict the SDF of a sphere centered at the origin with fixed radius. This helps to avoid being stuck in local minima with concave surfaces.

In addition, we also initialize the deformation network to predict a zero offset. This trick stabilizes the convergence of the network by first focusing the network capacity on improving the template before learning complex deformations.

13.1.4 Volume Renderer Training Losses

We now define the various shape-losses which are applied during the first stage of training. These are used to regularize the Morphable Volume Renderer, in order to avoid degenerate solutions and local minima.

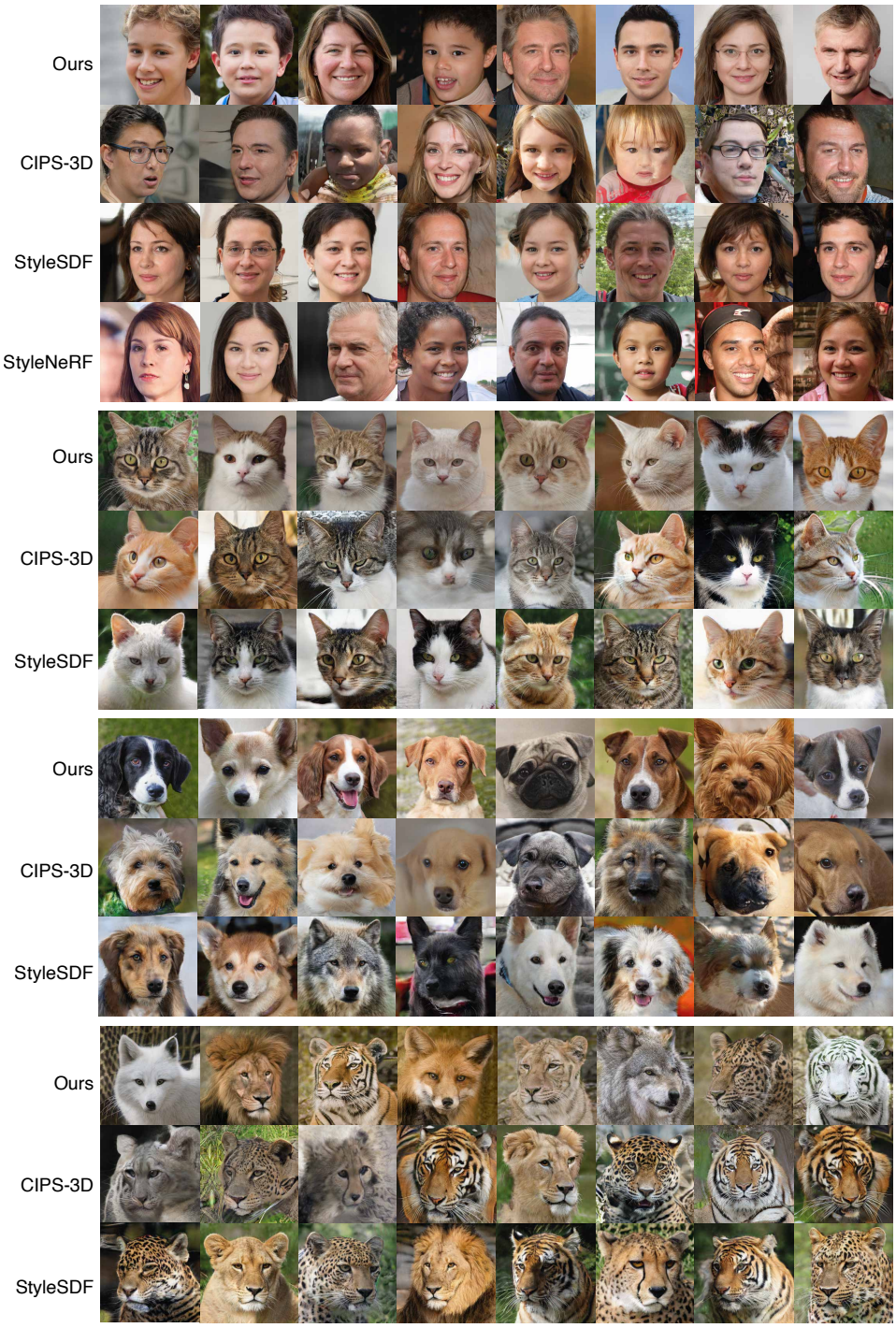

Figure 16: We match the visual synthesis quality of 3 state-of-the-art 3D-aware GANs, with the added benefit of controllability over multiple sources of variation. We compare with CIPS3D, StyleSDF and StyleNERF (FFHQ only).

Following Gropp et al. (2020); Wang et al. (2021a); Or-El et al. (2021), we use Eikonal loss to regularize the SDF d on the sampled points $\mathbf{p}$:

$$\mathcal{L}_{eik} = \mathbb{E}(\|\nabla d(\mathbf{p})\|_2 - 1)^2 \tag{5}$$

Following Or-El et al. (2021), we use Surface loss to avoid unnecessary surface crossings:

$$\mathcal{L}_{surf} = \mathbb{E}(\exp(-100|d(\mathbf{p})|)) \tag{6}$$

Hyperparameter	value
Optimizer	Adam
Gen LR	2e-5
Disc LR	2e-4
β	(0,0.9)
Loss weights	
λ_{surf}	0.05
λ_{eik}	0.1
λ_{bin}	1
λ_{cov}	1
Volume Rendering	
Points per ray	24

Table 3: Hyperparameters used to train the volume renderer for the first stage training

To facilitate training of our volume renderer to produce a realistic mask M, we include 2 mask-based losses following Xue et al. (2022) with minor modification. Binary loss (Eq. 7) encourages mask values close to 0 or 1 and Coverage loss (Eq. 8) encourages a mean mask value close to 0.75.

$$\mathcal{L}_{bin} = \mathbb{E}(\min(M, 1 - M)) \tag{7}$$

$$\mathcal{L}_{cov} = \mathbb{E}(|M - 0.75|) \tag{8}$$

The shape losses have hyperparameter loss weights λ_*. These weights, along with all other hyperparameter values used during training are shown in Table. 3.

Our full morphable volume renderer training objective is as follows:

$$\begin{aligned} \mathcal{L}_{vol} &= \mathcal{L}_{gen} + \lambda_{surf}\mathcal{L}_{surf} + \\ &\lambda_{eik}\mathcal{L}_{eik} + \lambda_{bin}\mathcal{L}_{bin} + \lambda_{cov}\mathcal{L}_{cov} \end{aligned} \tag{9}$$

13.2 Training the full pipeline

In the second phase, the full pipeline is trained but the volume renderer weights are kept frozen. For this stage, we rely on the default StyleGan2 (Karras et al., 2020b) training hyperparameters, using Adaptive Discriminator Augmentation (Karras et al., 2020a), and with path regularization set to 0. We set our alpha-consistency loss weight $\mathcal{L}_{alpha}$ to 5 for AFHQ categories, and 3 for FFHQ.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13970–13979, October 2021. 1, 12

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 5

Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)*, 41(1):1–26, 2022. 2, 3

Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. *CVPR*, 2022. 7

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 14

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 8

Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 11

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020a. 15

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020b. 15

Filippos Kokkinos and Iasonas Kokkinos. To the point: Correspondence-driven monocular 3d category reconstruction, 2021. 11

Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 452–461, 2020. 11

Feng Liu and Xiaoming Liu. Learning implicit functions for topology-varying dense 3d shape correspondence. In *In Proceeding of 2020 Conference on Neural Information Processing Systems*, Virtual, December 2020. 7

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, jan 2019. doi: 10.1109/taffc.2017.2740923. URL https://doi.org/10.1109%2Ftaffc.2017.2740923. 6, 8

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *arXiv e-prints*, pp. arXiv–2112, 2021. 13, 14

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 12

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 1, 3

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. URL https://doi.org/10.1007/s11263-010-0380-4. 3

Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. *arXiv preprint arXiv:2203.15926*, 2022. 2, 3

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021a. 14

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9168–9178, 2021b. 5

Ziyu Wang, Yu Deng, Jiaolong Yang, Jingyi Yu, and Xin Tong. Generative Deformable Radiance Fields for Disentangled Image Synthesis of Topology-Varying Objects. *Computer Graphics Forum*, 2022. 7

Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. *arXiv preprint arXiv:2203.14954*, 2022. 2, 3, 15

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018. 1, 3, 5

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 8