Learned Spatial Representations for Few-shot Talking-Head Synthesis

Moustafa Meshry Saksham Suri Larry S. Davis Abhinav Shrivastava

University of Maryland, College Park

Abstract

We propose a novel approach for few-shot talking-head synthesis. While recent works in neural talking heads have produced promising results, they can still produce images that do not preserve the identity of the subject in source images. We posit this is a result of the entangled representation of each subject in a single latent code that models 3D shape information, identity cues, colors, lighting and even background details. In contrast, we propose to factorize the representation of a subject into its spatial and style components. Our method generates a target frame in two steps. First, it predicts a discrete and dense spatial layout for the target image. Second, an image generator utilizes the predicted layout for spatial denormalization and synthesizes the target frame. We experimentally show that this disentangled representation leads to a significant improvement over previous methods, both quantitatively and qualitatively.

1. Introduction

We study the task of learning personalized head avatars in a low-shot setting, also known as *"neural talking heads"*. Given a single-shot or few-shot images of a source subject, and a driving sequence of facial landmarks, possibly derived from a different subject, the goal is to synthesize a photo-realistic video of the source subject, under the poses and expressions of the driving sequence. This task has a wide range of applications, including those in AR/VR, video conferencing, gaming, animated movie production and video compression in tele-communication.

Traditional graphics-based approaches to this task rely on a 3D face geometry and produce very high quality synthesis. However, they tend to focus on modeling the face area without the hair, and they learn a subject-specific model and cannot generalize to new subjects. In contrast, recent 2D-based approaches [1, 2, 3, 4] learn a subject-agnostic model that can animate unseen subjects given as few as a single image. Furthermore, since these works learn an implicit model and do not require an explicit geometric representation, they can synthesize the full head, including the hair, mouth interior, and even wearable accessories like glasses and earrings. This remarkable generalization ability however comes at the cost of low quality and poor identity preservation when compared to their 3D-based subject-specific counterparts. Bridging the quality gap between 2D-based subject-agnostic and 3D-based subject-specific approaches remains an open problem.

Figure 1: Our framework factorizes the image synthesis process into its spatial and style components. It predicts a discrete latent spatial layout for the target image, which is used to produce per-pixel style modulation parameters for the final synthesis.

Recent efforts in 2D-based approaches can be divided into two classes; *warping-based* and *direct synthesis*. As the name suggests, warping-based methods (*e.g.,* [2]) learn to warp the input image or a recovered canonical pose based on the motion of the driving sequence. While these methods achieve high realism, especially for static and rigid parts of the image, they tend to work well only for a limited range of motion, head rotation and dis-occlusion. On the other hand, direct synthesis approaches (*e.g.,* [1, 3, 4]) encode the source subject into a compressed latent code, and a generator decodes the latent code to synthesize the target pose. These approaches learn a prior over the compressed latent space, and can generate realistic results for a wider range of poses and head motion. However, they exhibit a noticeable identity gap between their output and the source subject.

We posit that the identity gap is caused by the entangled representation of the source subject in a single latent code. This compressed 1D latent encodes multi-view shape information, identity cues, as well as color information, lighting and background details. In order to synthesize a target view from a latent code, the generator needs to devise a complex function to decode the uni-dimensional latent into its corresponding 2D spatial information. We argue this not only consumes a large portion of the network capacity, but also

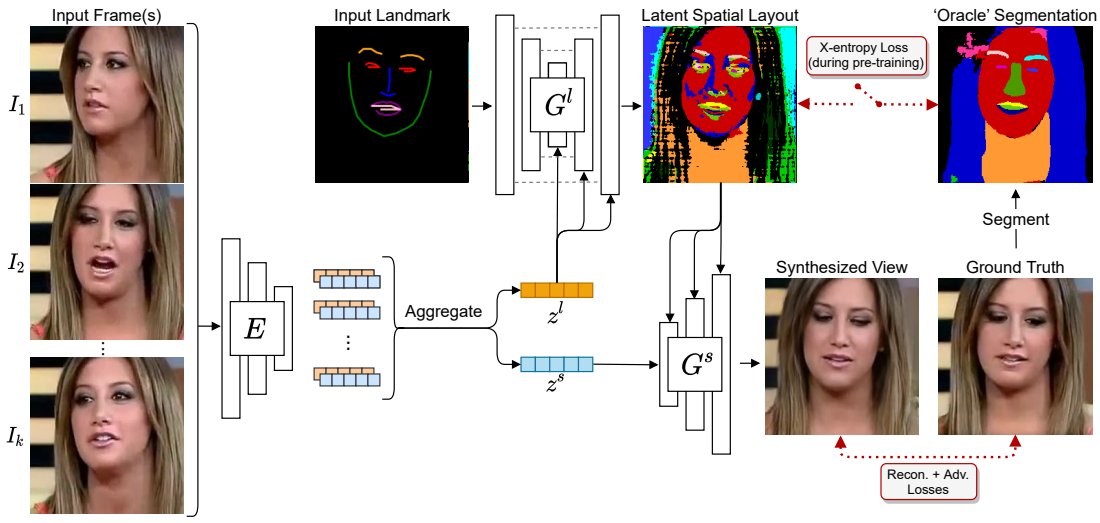

Figure 2: Overview of our training pipeline. The cross-entropy loss with the oracle segmentation is used during pre-training the layout predictor G^l, and then turned off during the full pipeline training.

limits the amount of information that can be encoded in the latent code.

To address this problem, we propose a two-step framework that decomposes the synthesis of a talking head into its spatial and style components. Our framework animates a source subject in two steps. First, it predicts a novel spatial layout of the subject under the target pose and expression. Then, it synthesizes the target frame conditioned on the predicted layout. This factorized representation yields the following key performance advantages.

Better subject-agnostic model performance. The performance of our subject-agnostic (also called *meta-learned*) model not only performs better than previous subject-agnostic state-of-the-art, but is also on-par with the subject-finetuned performance of previous works when there are only few source images available (*e.g.,* less than 10 images).

Better fine-tuned performance with less data. Fine-tuning our model for a specific subject requires significantly less data and fewer iterations than previous works, and yet achieves better performance. For example, we show that fine-tuning our model using 4-shot inputs outperforms previous state-of-the-art models fine-tuned using 32-shot inputs.

Robustness to pose variations. We show that our model is more robust against a wider range of poses and facial expressions, while still producing both realistic and identity-preserving results.

Improved identity preservation. Shape difference between the source and driving identities poses a challenge for identity preservation in reenacted results. The intermediate novel spatial representation learned by our model reduces the sensitivity towards such differences and better preserves the identity.

In summary, we make the following contributions:

- A novel approach that factorizes the *talking-head* synthesis process into its spatial and style components.
- A novel latent spatial representation that proves effective for few-shot novel view synthesis.
- We achieve state-of-the-art performance in both the single-shot and multi-shot settings, as well as in the meta-learned and subject-finetuned modes.

2. Related work

Existing approaches for realistic talking-head synthesis can be categorized into 3D-based and 2D-based.

3D-based methods. Such methods [5, 6, 7] utilize 3D geometric representations as a proxy to animate a target subject. Common geometric representations, such as 3D morphable models (3DMM) [8], only model the face area, and do not include challenging regions like the hair, eyes and mouth interior. Obtaining a detailed geometry of these regions is an expensive and challenging task. Therefore, such methods either cannot synthesize or perform poorly on those regions. Recent works [9, 10, 11] combine the traditional graphics pipeline with machine learning to better model the eye movement, mouth interior, or learn a better appearance model. However, they learn subject-specific models that do not generalize to new subjects. Other works [12, 13] take first steps to generalize to multiple subjects but they do not perform well on hair and other regions outside the face.

2D-based methods. These methods [1, 2, 3, 4, 14, 15, 16, 17, 18, 19] learn an implicit model of the head and do not require a proxy geometry. Therefore they can synthesize the full head including dynamic regions like the hair, eyes, and mouth interior. They can also model different wearable

accessories such as hats, glasses, and earrings. Early works build on top of CycleGAN [20] and learn subject-specific models [21, 22]. More recent works [1, 2, 3, 4, 18, 19] learn subject agnostic models that can animate unseen subjects given only a single or few-shot images. However, these methods lack in quality and identity preservation compared to the 3D-based subject-specific models. To bridge this performance gap, hybrid models [1, 3, 4] utilize a meta-learning phase that trains a subject-agnostic model on a large corpus of data, then an optional subject-specific fine-tuning phase is performed to improve the realism and restore the source identity. In this work, we improve the meta-learned performance to achieve state-of-the-art results without any subject-specific fine-tuning. While our model could still benefit from the optional fine-tuning phase to further refine the results, it requires significantly less data samples compared to previous works.

On another axis, 2D-based approaches can be categorized based on the synthesis technique into warping-based (*e.g.,* [2, 18, 23, 24]) and direct synthesis (*e.g.,* [1, 3, 4]). Warping-based approaches warp an input image [2, 18] or a recovered canonical pose [23] to synthesize novel poses. Warping results however tend to break when the target pose is far from that of the source image. Direct synthesis approaches utilize advances in Generative Adversarial Networks (GANs) [25] and Image-to-Image (I2I) translation [26] to generate novel poses. Compared to warping-based approaches, direct synthesis methods can realistically handle a wider range of poses and expressions. Concurrent to our work, Wang *et al*. [24] combine a warping-based approach with the power of GANs to achieve remarkable results.

Multi-modal Image-to-Image (I2I) translation. Several multi-modal I2I translation works feed a style latent code, either directly to the generator [27, 28, 29] or through adaptive instance normalization (AdaIN) [30, 31]. Recent state-of-the-art architectures [32, 33, 34] showed a significant improvement over traditional UNet [35] and encoder-decoder architectures, by generating per-pixel spatial denormalization (SPADE) parameters [32]. However, such architectures depend on the existence of accurate semantic segmentations or other dense spatial representations of the target image, hence limiting their usage in tasks where such dense representations do not exist. In this work, we learn to predict a latent dense layout to provide the spatial input to SPADE.

3. Method

Our approach factorizes the representation of a head avatar into spatial and style components. It breaks down the novel view head synthesis of a subject into two steps. First, a layout prediction network G^l translates facial landmarks for a target view into a dense spatial layout of the subject. Then, an image generator G^s synthesizes the final image conditioned on the predicted layout. We first give an overview of our pipeline in Section 3.1. Then, we explain how to pre-train the layout prediction network G^l to predict semantic segmentations of novel views in Section 3.2, followed by the full pipeline training in Section 3.3. Section 3.4 explains how the layout prediction network G^l transitions from predicting semantic maps to learning a more powerful latent spatial representation. And finally, we discuss how to learn a personalized head avatar through an optional subject-specific fine-tuning stage in Section 3.5.

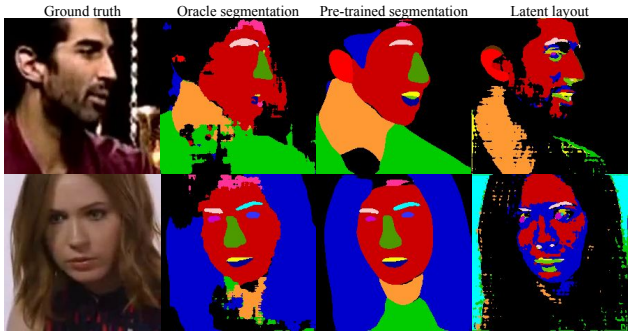

Figure 3: Layout pre-training predicts meaningful segmentation maps despite the noisy oracle segmentations. Our latent spatial representation encodes more information than traditional segmentations.

3.1. Overview

Given K-shot inputs $\{I_1 \dots I_K\}$ of a source subject, a two-headed encoder $E = \{E^l, E^s\}$ processes the inputs and generates K layout latents $\{z_i^l\}$ and K style latents $\{z_i^s\}$ for $i \in \{1 \dots K\}$. The K latents are then averaged to get an aggregated layout latent $z^l = \frac{1}{K}\sum_{i=1}^{K} z_i^l$ and style latent $z^s = \frac{1}{K}\sum_{i=1}^{K} z_i^s$. Averaging the K latents cancels out view-specific information and transient occluders, and maintains implicit 3D information like the head and hair shape for the layout latent, and color and lighting information for the style latent. We have two generators: a layout predictor network G^l and an image generator G^s. The layout predictor takes as input the facial landmarks for a target view x_t and the layout latent z^l and generates a spatial one-hot layout $y_t^l = G^l(x_t, z^l)$, such as a semantic map, for the target view. The image generator G^s processes the style latent z^s and utilizes spatial denormalization layers (SPADE [32]), conditioned on the predicted layout y^l, to synthesize the final image $\hat{I} = G^s(y^l, z^s)$. An overview of our framework is shown in Figure 2.

3.2. Layout prediction pre-training

Training the above pipeline end-to-end without any supervision or constraints on the predicted layouts results in a degenerate solution, where the spatial layouts and their corresponding spatial denormalization are completely ignored. All spatial and style information are thus encoded into and

decoded from the style latent z^s, which results in a poor performance. Therefore, we opted to pre-train the layout prediction network to predict a plausible semantic segmentation of a target view, given the input facial landmarks x_t and the layout latent z^l. To supervise this training, we use an off-the-shelf face segmentation network [36] as an oracle to segment the target image I_t into a semantic map S_t, and we apply a cross-entropy loss (X-ent) between the oracle segmentation S_t and our predicted segmentation $y_t^l = G^l(x_t, z^l)$. We observe that the obtained oracle segmentations are very noisy and have poor quality (*e.g.,* Figure 3). This is caused by the domain gap, in terms of image resolution and the distribution of head poses, between the datasets used to train the oracle segmentation network [36], and in-the-wild videos of talking heads. Thus, to regularize the segmentation prediction training, we use a mutli-task pre-training strategy where the layout prediction network predicts an extra RGB reconstruction R_t of the target image I_t, which is used as a secondary supervisory signal. Specifically, we have

$$y_t^l, R_t = G^l(x_t, z^l), \qquad z^l = \frac{1}{K}\sum_{i=1}^{K} E^l(I_i) \tag{1}$$

And the objective for the pre-training is

$$\mathcal{L}_{\text{seg}} = \text{X-ent}(y_t^l, S_t) + \lambda_R \mathcal{L}_R(R_t, I_t) \tag{2}$$

where $\mathcal{L}_R$ is a perceptual reconstruction loss, and λ_R is a relative weighting term which is set to a low value.

3.3. Full pipeline training

Once the layout predictor network has been pre-trained to predict semantic segmentations, we plug it into our full pipeline. The predicted segmentation is fed as the spatial input to a SPADE image generator G^s that synthesizes the final image as

$$\hat{I} = G^s(G^l(x_t, z^l), z^s), \qquad z^s = \frac{1}{K}\sum_{i=1}^{K} E^s(I_i) \tag{3}$$

We observe that the SPADE generator quickly utilizes the input spatial segmentations to resolve spatial ambiguities, and we no longer fall into a degenerate solution where the spatial input is ignored.

Our full pipeline, comprising the layout and style encoders $\{E^l, E^s\}$, the layout predictor G^l and the image generator G^s, is optimized to minimize three losses; a reconstruction loss $\mathcal{L}_{\text{rec}}$, an adversarial loss $\mathcal{L}_{\text{adv}}$, and a latent regularization loss $\mathcal{L}_{L2}$.

For the reconstruction loss $\mathcal{L}_{\text{rec}}$, we employ a perceptual loss [37] based on both the VGG19 [38] and VGGFace [39] networks, as well as an $L1$ loss. While the VGG19-based perceptual loss is a standard reconstruction loss, we follow Zakharov *et al*. [1] and utilize a VGGFace-based perceptual loss to promote identity preservation. We also use an $L1$ loss to better preserve color transfer between the synthesized and ground truth images.

The adversarial loss, $\mathcal{L}_{\text{adv}}$, encourages the output to be photo-realistic. To achieve that, a discriminator network D is trained to discriminate between real and fake images, while the generator network, G^s aims to fool the discriminator by bringing the output closer to the manifold of real images. We borrow the architecture of the discriminator network D from [40] and use a non-saturating logistic loss with gradient penalty [41]. Finally, we impose an $L2$ regularization on the learned latent codes to encourage compactness of the latent space. The full training objective is given by

$$\begin{aligned}\min \mathcal{L}(\hat{I}_t, I_t, z^l, z^s | E^l, E^s, G^l, G^s, D) = \mathcal{L}_{\text{rec}}(\hat{I}_t, I_t) + \\ \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(\hat{I}_t, I_t) + \lambda_{L2}\left(\|z^l\|^2 + \|z^s\|^2\right)\end{aligned} \tag{4}$$

where $\lambda_{\text{rec}}, \lambda_{L2}$ determine the relative weights between the loss terms.

3.4. Learning a latent spatial representation

Spatial denormalization (SPADE) generates per-pixel denormalization parameters by feeding a dense spatial input through a small convolutional subnetwork. While SPADE [32] originally uses semantic maps as input, we explore learning a latent spatial representation that better suits the image synthesis task at hand. To do this, we turn off the cross-entropy loss so as to give the layout predictor G^l the freedom to diverge from predicting traditional semantic segmentations and learn other latent representations that better optimize the few-shot novel view synthesis objective. The layout predictor is thus supervised only by the training objective of Eqn. 4. Figure 3 shows examples of the learned latent layouts. Although they might look less interpretable than traditional semantic maps, they seem to encode more information and capture accurate details.

3.5. Subject fine-tuning

Training our full pipeline learns a powerful subject-agnostic model that produces high quality and identity-preserving synthesis. Optionally, we can learn a personalized head avatar to further refine the results for a given subject. To do this, we follow [1, 3, 4] and fine-tune the subject-agnostic model (also called *meta-learned* model) using the few-shot inputs of the source identity. Specifically, we compute the layout and style embeddings $\{z^l, z^s\}$ and fine-tune the weights of the layout and image generators $\{G^l, G^s\}$, as well as the discriminator, D, by reconstructing the set of few-shot inputs, and optimizing the same training objective of Eqn. 4. We observe that subject fine-tuning restores high-frequency components and improves background reconstruction when compared to the meta-learned outputs.

Figure 4: Qualitative comparison in the single-shot setting. We show three sets of examples representing low, medium and high variance between the source and target poses. Our method is more robust to pose variations than the baselines.

4. Experimental evaluation

Implementation details. Please, refer to the supp. material for networks architecture, hyper-parameters and training details. Our code will be publicly released.

Dataset. We perform our evaluation using the VoxCeleb [42] dataset, which is a large-scale in-the-wild video dataset. The train set contains over a million clips from 145,569 videos of 5,994 different identities. The test set contains new identities that are not part of the training. We use the test subset released by Zakharov *et al*. [1], which contains a total of 1,600 frames from videos of 50 subjects. For self-reenactment scenarios, the input few-shots and the driving sequence do not overlap. We obtain the facial landmarks for sampled frames using an off-the-shelf facial landmarks detector [43].

Table 1: Quantitative comparison in the single-shot setting.

Method	PSNR↑	SSIM↑	LPIPS↓	ID-SIM↑	NMKE↓	FID↓
X2Face [23]	15.50	0.466	0.346	0.691	0.333	98.58
Bi-layer [4]	–	–	–	0.721	0.236	130.58
FSTH [1]	16.92	0.597	0.263	0.836	0.049	53.07
LPD [3]	–	–	–	0.837	0.070	48.48
FOMM [2]	**18.20**	**0.635**	0.236	0.869	0.061	56.10
Ours	17.37	0.605	**0.232**	**0.886**	**0.041**	**45.69**

Baselines. We compare our method to the following baselines: X2Face [23], FSTH [1], FOMM [2], Latent Pose Descriptor (LPD) [3], and Bi-layer [4]. We use the released pre-trained models provided by the authors for all baselines, except for FSTH [1] where we use the authors' provided outputs, as their code and models were not released. Since some baselines only accept single-shot inputs (*e.g.,* FOMM and Bi-layer), we divide our evaluation into a single-shot setting, where we compare to all the baselines, and a multi-shot setting, where we only compare against the few-shot baselines. Since the LPD [3] and Bi-layer [4] baselines do not predict the background and re-crop the input/output frames, we subtract the background and compare with their corresponding cropped ground truths for quantitative analysis. We also exclude those two baselines from frame reconstruction evaluation since their output does not align with the rest of the methods.

Metrics. We evaluate all models along five axes.

- Reconstruction fidelity using the peak signal-to-noise ratio (*PSNR*) and structural similarity (*SSIM*) [44] metrics.
- Perceptual similarity between the output and the ground truth using the *AlexNet*-based *LPIPS* metric [45].
- Identity preservation (*ID-SIM*) using the cosine similarity between face embeddings from a face recognition network [39].
- Normalized Mean Keypoint Error (*NMKE*), which measures the pose error between the synthesized and ground truth images as computed in [3, 4].
- Perceptual quality of the output using the Frechet-Inception Distance (*FID*) metric [46].

4.1. Single-shot comparative evaluation

Table 1 shows a quantitative comparison with the baselines in the single-shot setting. Our method outperforms all baselines in perceptual reconstruction (LPIPS), identity preservation (ID-SIM), pose matching (NMKE) and visual quality (FID). However, FOMM scores better in the standard reconstruction metrics (PSNR and SSIM). We argue this is intrinsic to their method due to its warping-based nature, which accurately captures the background and other static regions, and thus gives low reconstruction error even in the presence of clear artifacts. Furthermore, while FOMM cannot utilize more input frames to its advantage, our method's performance improves with multi-shot inputs to significantly surpass FOMM in all metrics (see supp. material for the quantitative numbers).

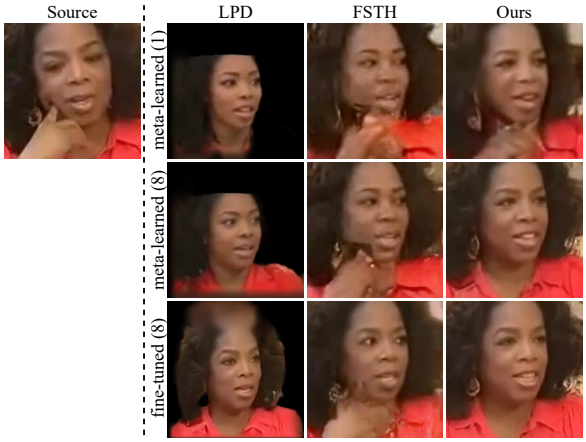

Figure 5: A qualitative comparison showing the effect of increasing the K-shot inputs and applying subject fine-tuning.

Figure 4 shows qualitative results from three groups representing low, medium and high variance between the input and target poses. We observe that all methods perform well when the target pose is similar to that of the input shot. LPD produces sharp results within the low-medium pose variation, but shows blurry artifacts within the face and eyes in the case of high pose variance. FSTH shows a clear identity gap. FOMM accurately matches the background and shows highly realistic results when the pose variance is low, but shows a clear identity gap and visible artifacts when the target pose is far from the source image. Our method is more robust against pose variation, yielding realistic results while preserving the source identity.

4.2. Multi-shot comparative evaluation

Here, we focus on the effect of increasing the number of K-shot inputs, and the effect of subject-specific fine-tuning using the K-shot inputs. Figure 6 plots the ID-SIM, NMKE and FID performance metrics as we increase the number of K-shots. We observe that the pose reconstruction performance (NMKE) is mainly dictated by the approach itself, rather than the number of K-shots or whether the models are fine-tuned or not. For example, the *meta-learning* performance of FSTH with $K = 1$ is better than the *fine-tuned* LPD model with $K = 32$. Similarly, the *single-shot meta-learning* performance of our method is better than the *fine-tuned* baselines at $K = 32$.

For the ID-SIM and FID metrics, the *meta-learning* performance of our model is not only superior to that of the baselines, but it is also on-par with the *fine-tuned* baselines for $K \leq 8$. However, as K is increased to 32, the *fine-tuned* baselines eventually outperform our *meta-learned* model. Another very important advantage to our approach is that it achieves better performance with significantly less data.

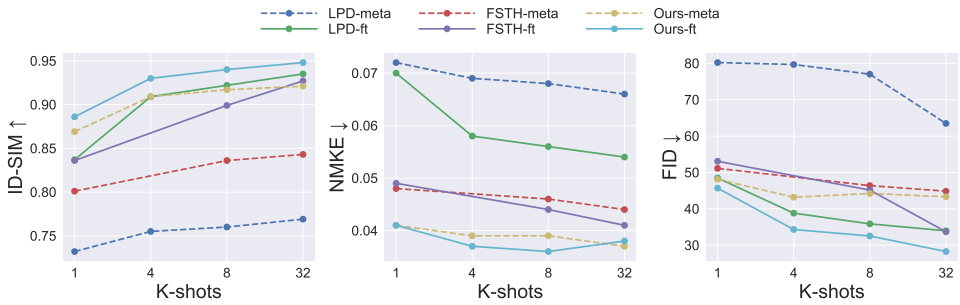

Figure 6: Quantitative comparison with the few-shot baselines, showing the effect of both increasing the K-shot inputs and subject-specific fine-tuning. Dotted and solid lines represent the meta-learned and fine-tuned models respectively.

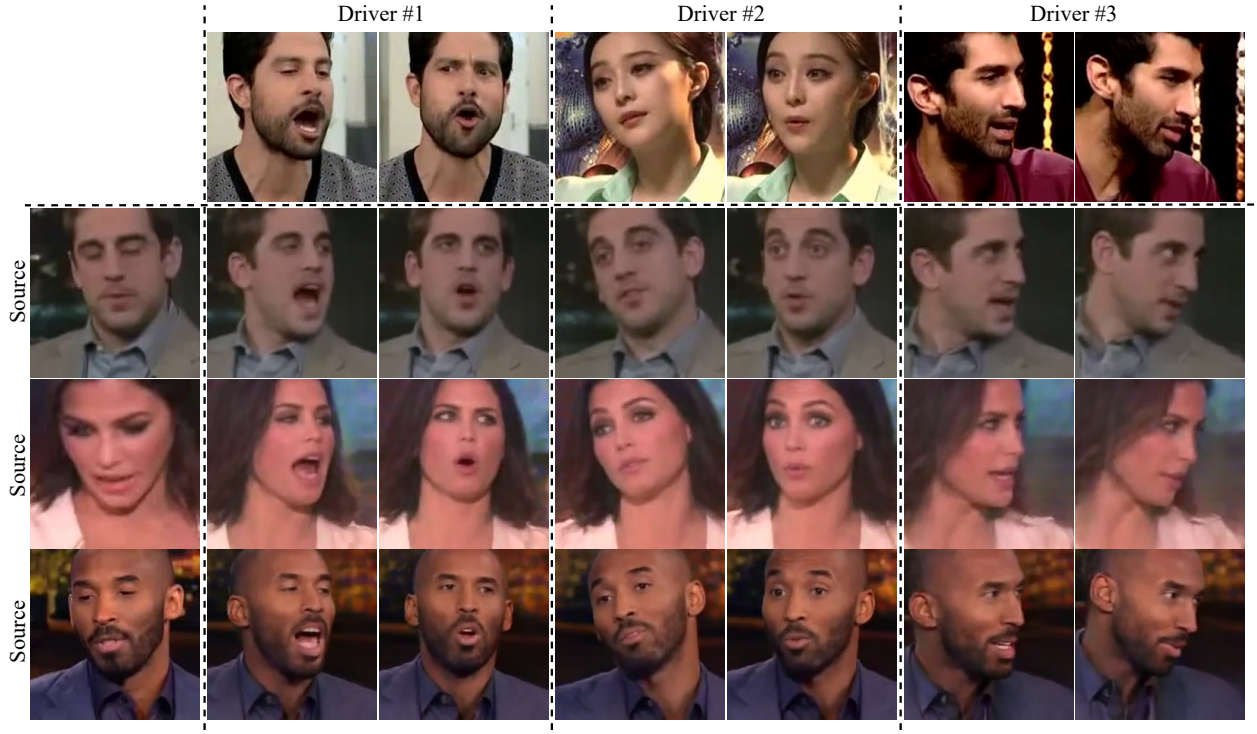

Figure 7: Cross-subject reenactment with different driving identities. Results are shown for our *meta-learned* model without any fine-tuning, and using 32-shot inputs.

For example, fine-tuning our model with just $K = 4$ outperforms the fine-tuned baselines at $K = 32$. Since fine-tuning on more data requires more training iterations and thus more time, our method spends much less time fine-tuning on fewer data samples, and yet achieves similar or better results. We observe similar behavior with other metrics (PSNR, SSIM and LPIPS). Please, refer to the supp. material for full results.

Figure 5 visualizes the effect of both increasing K and subject fine-tuning. Our method preserves the source identity without any fine-tuning, even with a single-shot input. On the other hand, the baselines only restore the source identity after the subject-specific fine-tuning. Our method also shows the most improvement, in terms of realism and better identity match, when increasing the number of K-shot inputs. For example, our method successfully filters out the subject hand occluding the face in the single-shot input.

4.3. Cross-subject reenactment

Cross-subject reenactment poses a challenge, especially for landmark-driven approaches. The shape difference between facial landmarks of the source and driver identities could lead to a noticeable identity gap in the reenacted results. The intermediate spatial representation learned by our method helps reduce this problem and leads to good identity preservation of the source subject regardless of the driver identity. Figure 7 shows sample reenactment results using different driver identities. To demonstrate the effectiveness of our disentangled representation, we avoid any subject fine-tuning and show the results of our meta-learned model with 32-shot inputs. The source identity is well-preserved among challenging facial expressions and different views covering both the left and right sides of the face.

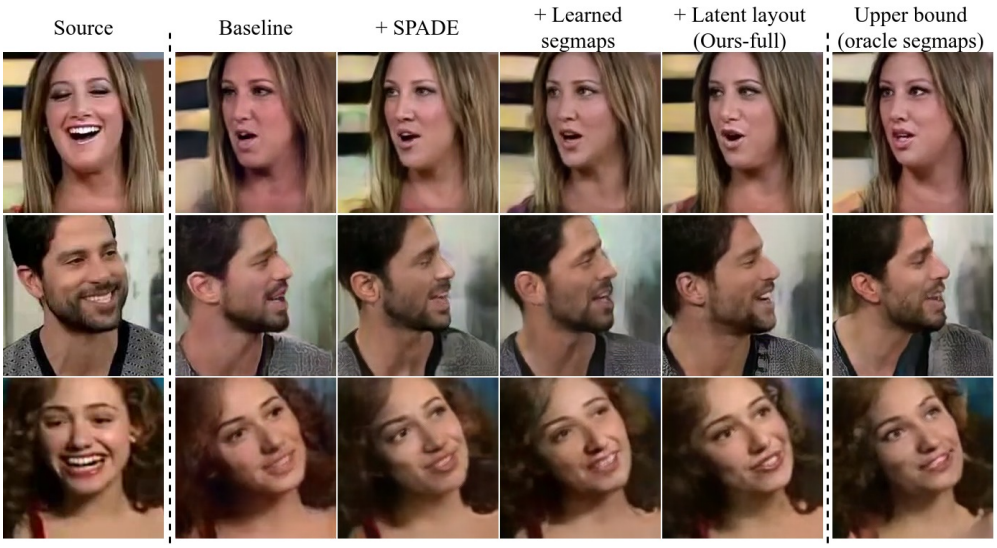

Figure 8: Examples from the ablation study. Results shown are for the meta-learned models with a single-shot input (source).

Table 2: Ablation study of our approach. +SPADE replaces the UNet generator with SPADE. +Learned seg. maps conditions the generator on learned segmentations. +Latent layout learns a latent spatial representation. The upper bound gets to cheat and uses the ground truth segmentations.

Method	PSNR↑	SSIM↑	LPIPS↓	ID-SIM↑	NMKE↓	FID↓
Baseline	17.00	0.574	0.274	0.837	0.044	67.19
+ SPADE	16.94	0.575	0.268	0.834	0.043	56.00
+ learned seg. maps	16.94	0.578	0.265	0.828	**0.042**	62.78
+ latent layout (**ours**)	**17.22**	**0.592**	**0.247**	**0.860**	0.042	**54.40**
Upper bound	18.21	0.629	0.219	0.867	0.039	48.06

4.4. Ablation study

We evaluate the contribution of different components of our proposed approach. All ablation experiments are trained with the same hyper-parameters and for the same number of epochs, and are evaluated in the *single-shot* setting with *no fine-tuning*. We report the results in Table 2. The baseline model has the same setup as FSTH [1], where a UNet generator with AdaIN layers [30] translates the input landmarks into the target image. Next, we replace the UNet architecture with a SPADE generator [32] conditioned on the facial landmarks (+SPADE). This improved the FID, but other metrics remained around the same. We hypothesize this is due to using sparse landmarks as the spatial input, while SPADE needs dense spatial inputs to generate the per-pixel denormalization parameters. To validate our hypothesis, we conducted an experiment as an *upper bound*, where we get to cheat and segment the ground truth target image using an off-the-shelf face segmentation network [36] (*i.e.* oracle), and we use these oracle segmentations as the spatial input to SPADE. Even though the oracle segmentations are noisy (*e.g.,* Figure 3), this still resulted in a significant boost in all metrics, proving that the SPADE generator could benefit from dense spatial inputs. Therefore, we trained a layout prediction network to predict a plausible semantic segmentation for the target pose (+Learned seg. maps). This surprisingly produced mixed results and even caused a drop in the ID-SIM and FID scores. We posit this is because the noisy oracle segmentations do not provide a consistent supervisory signal, which causes the learned segmentations to miss important shape cues (*e.g.,* the correct face shape), as well as overfit common errors in the oracle segmentations as the training progresses. Finally, removing the supervision on the predicted layouts and learning a latent spatial representations (+Latent layouts) resulted in a reasonable performance improvement over all metrics. We also show a qualitative comparison for the ablation study in Figure 8. We observe that the qualitative results of the upper bound experiment (using the oracle/ground-truth segmentation) exhibits artifacts caused by errors in the oracle segmentation. The results of our method with the learned latent layouts looks qualitatively better, with no clear artifacts, despite having worse quantitative metrics than the upper bound experiment.

5. Conclusion

We proposed a novel approach for talking-head synthesis. Our model learns a novel latent spatial representation that proves effective for our task. We improve the performance of both subject-agnostic and subject-finetuned models while requiring significantly less data samples. The learned latent spatial representation provides robustness against a wide range of poses and expressions, and results in better identity preservation, especially for the cross-subject reenactment scenarios.

Acknowledgements. We would like to thank the members of the Perception and Intelligence (PI) Lab for their helpful feedback. This project was partially funded by DARPA SemaFor (HR001119S0085), DARPA MediFor (FA87501620191) and DARPA SAIL-ON (W911NF2020009) programs.

References

[1] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9459–9468, 2019. 1, 2, 3, 4, 5, 6, 8

[2] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, December 2019. 1, 2, 3, 6

[3] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13786–13795, 2020. 1, 2, 3, 4, 6

[4] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Eur. Conf. Comput. Vis.*, August 2020. 1, 2, 3, 4, 6

[5] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015. 2

[6] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2387–2395, 2016. 2

[7] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4):1–13, 2017. 2

[8] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, pages 187–194, 1999. 2

[9] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. In *Proc. SIGGRAPH*, 2018. 2

[10] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 2019. 2

[11] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. *arXiv preprint arXiv:2012.03065*, 2020. 2

[12] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM Trans. Graph.*, 37(6):1–12, 2018. 2

[13] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38(4):1–14, 2019. 2

[14] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7832–7841, 2019. 2

[15] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Int. Conf. Comput. Vis.*, pages 7184–7193, 2019. 2

[16] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Eur. Conf. Comput. Vis.*, pages 818–833, 2018. 2

[17] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *AAAI*, volume 34, pages 10861–10868, 2020. 2

[18] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, volume 34, pages 10893–10900, 2020. 2, 3

[19] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 2, 3

[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2223–2232, 2017. 3

[21] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Eur. Conf. Comput. Vis.*, pages 119–135, 2018. 3

[22] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Eur. Conf. Comput. Vis.*, pages 603–619, 2018. 3

[23] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Eur. Conf. Comput. Vis.*, pages 670–686, 2018. 3, 6

[24] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *arXiv preprint arXiv:2011.15126*, 2020. 3

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, pages 2672–2680, 2014. 3

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[27] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Adv. Neural Inform. Process. Syst.*, 2017. 3

[28] Moustafa Meshry, Saksham Suri, Larry S. Davis, and Abhinav Shrivastava. Step: Style-based encoder pre-training for multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[29] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European conference on computer vision*, pages 318–335. Springer, 2016. 3

[30] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Int. Conf. Comput. Vis.*, pages 1501–1510, 2017. 3, 8

[31] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Eur. Conf. Comput. Vis.*, 2018. 3

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3, 4, 8

[33] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 3

[34] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5104–5113, 2020. 3

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[36] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 8

[37] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, pages 694–711. Springer, 2016. 4

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Int. Conf. Learn. Represent.*, 2015. 4

[39] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *Brit. Mach. Vis. Conf.*, 2015. 4, 6

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. 4

[41] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 4

[42] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*, 2018. 5

[43] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Int. Conf. Comput. Vis.*, pages 1021–1030, 2017. 5

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 6

[46] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *neurips*, 2017. 6