

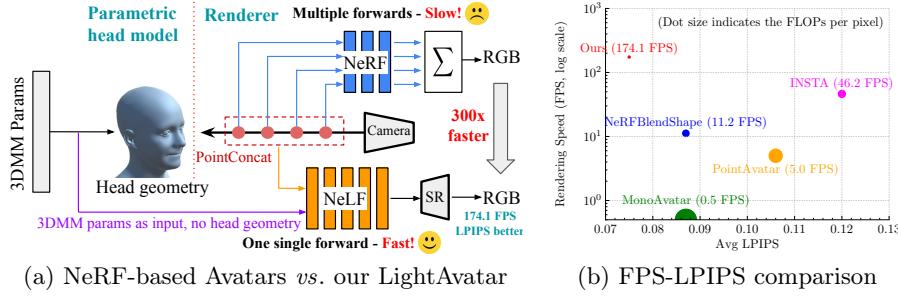
# LightAvatar: Efficient Head Avatar as Dynamic Neural Light Field

Huan Wang<sup>1,2,†</sup>, Feitong Tan<sup>2</sup>, Ziqian Bai<sup>2,3</sup>, Yinda Zhang<sup>2</sup>, Shichen Liu<sup>2</sup>, Qiangeng Xu<sup>2</sup>, Menglei Chai<sup>2</sup>, Anish Prabhu<sup>2</sup>, Rohit Pandey<sup>2</sup>, Sean Fanello<sup>2</sup>, Zeng Huang<sup>2</sup>, and Yun Fu<sup>1</sup>

<sup>1</sup> Northeastern University, USA

<sup>2</sup> Google, USA

<sup>3</sup> Simon Fraser University, Canada



**Fig. 1:** (a) Overview comparison between existing neural head avatars (top) *vs.* our LightAvatar (down) – a brand-new framework to build efficient 3D head avatars based on neural light field. LightAvatar features a simple and uniform design, which takes expression code and camera pose as input, renders the RGB via a single network forward pass, running at **174.1 FPS** (on a RTX3090 GPU) with image quality improved. (b) FPS and LPIPS comparison of recent top-performing (fast) avatars. Our method achieves much faster rendering speed with better LPIPS than the counterparts.

**Abstract.** Recent works have shown that neural radiance fields (NeRFs) on top of parametric models have reached SOTA quality to build photo-realistic head avatars from a monocular video. However, one major limitation of the NeRF-based avatars is the slow rendering speed due to the dense point sampling of NeRF, preventing them from broader utility on resource-constrained devices. We introduce **LightAvatar**, the *first* head avatar model based on neural *light* fields (NeLFs). LightAvatar renders an image from 3DMM parameters and a camera pose via a single network forward pass, without using mesh or volume rendering. The proposed approach, while being conceptually appealing, poses a significant challenge towards real-time efficiency and training stability. To resolve them, we introduce dedicated network designs to obtain proper representations for the NeLF model and maintain a low FLOPs budget. Meanwhile, we tap into a distillation-based training strategy that uses a pretrained avatar model as teacher to synthesize abundant pseudo data for training. A

<sup>†</sup>Work done when Huan was an intern at Google.

Corresponding author: Huan Wang, [huan.wang.cool@gmail.com](mailto:huan.wang.cool@gmail.com).

warping field network is introduced to correct the fitting error in the real data so that the model can learn better. Extensive experiments suggest that our method can achieve new SOTA image quality quantitatively or qualitatively, while being significantly faster than the counterparts, reporting **174.1 FPS** (512×512 resolution) on a consumer-grade GPU (RTX3090) with no customized optimization.

## 1 Introduction

Digitalizing a human is a long-standing problem in computer vision [11], which has recently attracted increasing attention due to its massive potential in AR/VR. One of the most prominent task is to build face or head avatars, which can generate vivid and photo-realistic appearance of the users upon controlling signals, *e.g.*, from parametric models [19, 42]. Recent advances in implicit neural representation (INR) [18, 48, 54] has motivated neural radiance field (NeRF) [50], which has been demonstrated as an enormous success in building controllable photo-realistic 3D head avatars [6, 20, 83].

Despite the remarkable rendering quality, existing NeRF-based avatars typically suffers from two issues. First, the rendering is computationally expensive, which is an issue arising from the neural volumetric rendering backbone due to multiple shading operations along each pixel ray. Although quite many works have attempted to accelerate *static* NeRFs [14, 17, 22, 57, 58, 72, 78], it is non-trivial to extend them to *dynamic* head avatars. Second, most of the high-quality avatars build upon explicit 3DMM [11, 19] geometry, *e.g.*, to derive a 3D deformation field [2, 26, 85], or directly anchor local radiance fields [6]. While delivering remarkable controllability and stability, these methods tend to perform poor when 3DMM geometry is over-simplified or missing.

In this paper, we present *LightAvatar*, a novel 3D head avatar model that renders high-quality images efficiently without leveraging an explicit geometry. Our model is fundamentally different from existing approaches in the rendering backbone (see Fig. 2), where we leverage a neural *light* field (NeLF) instead of a neural *radiance* field (NeRF) as scene representation. The rendering of light fields amounts to *a single* neural network forward pass *vs.* hundreds of network forward passes in NeRF [50]. The network efficiently renders a target image by a single forward pass, with 3DMM parameters and camera pose as input and transformed by our dedicatedly designed sub-networks into better representations for the NeLF backbone. To further speed up the rendering, we introduce an image *super-resolution (SR)* [43, 77, 81] module after the NeLF backbone in our pipeline, which enables us to feed a low-resolution input into the network while obtain a high-resolution output. As a result, the rendering is substantially faster than the other counterparts (174.1 FPS, see Tab. 3).

One downside of removing the dependency on explicit 3DMM geometry, however, is less training stability due to the missing of strong priors, especially when training from a monocular video. To overcome this, we tap into the recent advances of *knowledge distillation* [12, 29] in efficient neural rendering [14, 72, 79].

Specifically, we employ a pretrained avatar model to synthesize abundant pseudo data, and distill a LightAvatar model from them. To prevent the performance from being capped by the teacher model, we train jointly on both the pseudo and real data. While perfect 3DMM fitting is guaranteed on the pseudo data, this is not true on the real data. We observed that naively adding real data may even hurt the performance. To account for the fitting error in the real data, we introduce a *warping field network* to mitigate the fitting error in the real data, resulting in improved overall quality. Contributions of this work are:

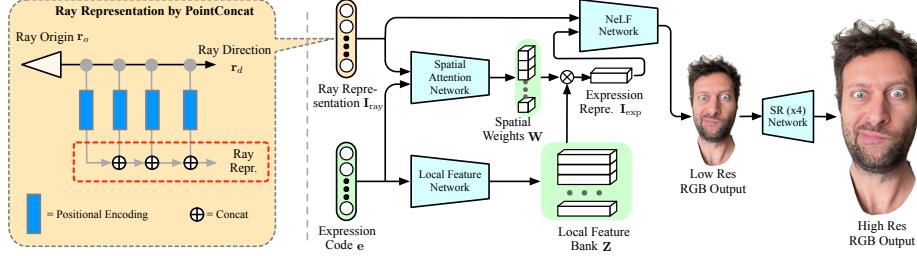
- We introduce *LightAvatar*, the *first* head avatar based on neural light fields (NeLFs) that does not rely on explicit meshes or volume rendering. This novel approach results in a simple and efficient pipeline.
- The method features several dedicated network designs: (1) The expression representation produced by our model outperforms the common baseline solution of using raw expressions as NeLF model input. (2) Importantly, we introduce an SR module that significantly improves the inference efficiency.
- We present a distillation-based training strategy with a warping field network for correcting fitting error to learn effectively on pseudo data and real data.
- Extensive empirical results and analyses on multiple subjects show our LightAvatar can achieve *consistently better* image quality (see Tab. 1, Tab. 2, Fig. 3, Fig. 4) than the counterparts while running at **174.1 FPS** on a RTX3090 GPU with no customized optimization (see Tab. 3).

## 2 Related Work

### 2.1 Monocular 3D Head Avatars and Fast Avatars

**Monocular 3D Head Avatars.** It has been a long challenge to reconstruct 3D head avatars from a monocular video. Leveraging the low-dimensional priors when modeling human heads (*e.g.*, 3DMM [11, 19]), many previous works model the geometry and texture *explicitly*, such as [13, 23, 24, 30, 31, 38, 68–70, 74]. By fitting a morphable model to a given subject through traditional optimization [86] or neural-based techniques [4, 5, 15, 66, 76], they employ shared mesh topology and texture parameterization, enabling subsequent animation and manipulation. Nevertheless, these approaches often face challenges due to their restricted representation ability, especially in modeling subtle details and components such as hair and accessories that fall outside the model parameters.

The rapid advancement of implicit neural representations (INR), represented by NeRF [7, 8, 50] has recently popularized the implicit modeling of avatars, thanks to its superior rendering quality and the ability to comprehensively represent the entire head. For instance, NerFACE [20] directly conditions neural radiance fields with 3DMM expression codes to achieve dynamic avatars. RigNeRF [2], on the other hand, employs expression conditioning in the canonical space, defined by a 3DMM-guided warping field. IMAvatar [83] endeavors to learn blendshapes and skinning fields to represent deformations dependent on avatar expressions and poses. Additionally, MonoAvatar [6] focuses on learning



**Fig. 2:** Overview of our LightAvatar. The method consists of four trainable networks (*spatial attention network*, *local feature network*, *NeLF network*, and *SR network*). (1) Given an expression code, the local feature network transforms it to a *local feature bank*, which stores the features of different local head regions. (2) Given a specific ray and the expression code, the spatial attention network outputs a vector of *spatial attention weights* to query the local feature bank to obtain the expression representation for that ray. (3) Then, given the ray and expression representation as input, the NeLF model predicts the desired (low-resolution) RGB. (4) Finally, the SR network generates a high-resolution image with the low-resolution image as input. Notably, LightAvatar predicts the target RGB via a single network forward, thus enabling fast rendering.

local features anchored on the 3DMM geometry, allowing these features to be deformed by 3DMM animation and interpolated within the 3D volume to facilitate avatar animation. Despite the encouraging progress, a primary challenge remains in achieving precise control over motions and expressions, and improving the efficiency during training and inference [46, 84, 85].

**Fast Avatars.** Many recent papers have tried to improve the efficiency of head avatars. INSTA [85] uses neural graphics primitives [51] embedded on a parametric face model, achieving fast reconstruction in less than 10 minutes and interactive rendering. NeRFBleShape [21] adopts multi-level voxel fields as the bases citing the idea of Instant-NGP [51], where each voxel is modeled by a lightweight NeRF, enabling fast training and rendering. FLARE [10] presents a relightable mesh-based avatar where materials and lighting are disentangled by different MLPs. Hash-grid encoding [51], neural split-sum approximation [36], and differentiable rasterization [40] are used for fast training and rendering. Using neither mesh nor neural implicit representations, PointAvatar [84] employs a deformable point-cloud-based representation, which can be rendered efficiently with standard differentiable rasterizer. Despite their improved rendering efficiency, the visual quality of most of these methods is yet to be satisfactory according to our empirical studies (Fig. 4). In this work, we aim to deliver an avatar of both fast rendering *and* high quality.

## 2.2 Efficient Neural Rendering

**NeRF Inference Acceleration.** While neural rendering methods [67], particularly NeRFs [7, 8, 50], offer superior quality in synthesizing novel views, their computational demands during inference often pose challenges for real-time sce-

narios and resource-constrained devices. Efforts to enhance the efficiency of general NeRF model rendering have primarily pursued three directions. Some methods [22, 28, 65, 78] optimize the rendering by employing precomputation, shifting from MLP forward passes to table lookup. Other approaches [16, 51, 57, 58] remap the entire NeRF scene into more efficient spatial structures that can be parallelized, achieving consistent speedups. Finally, the rendering speed can be boosted by reducing the number of samples per camera ray [17, 44, 52], either within the original NeRF architecture or by embracing alternative frameworks such as neural light fields (NeLFs) [14, 27, 45, 72].

**Neural Light Fields** (NeLFs) have emerged as a promising solution for rapid neural rendering thanks to their distinct advantage of requiring only a single forward pass, in contrast to numerous samples per ray in NeRF. While light fields have been studied [25, 41] as efficient scene representations in real-time image-based rendering in computer graphics, their recent adaptation to neural networks [9, 35, 49] enables efficient implicit scene modeling. Among them, light field networks [63], light field neural rendering [64], RSEN [3], NeuLF [45], R2L [72], MobileR2L [14], and LightSpeed [27] have proposed various strategies for efficient NeLF-based scene representations. Our work aligns with this trend by modeling neural avatars as NeLF instead of NeRF to significantly accelerate inference. However, compared to existing methods that primarily address static scenes, training dynamic NeLF on animatable avatars presents greater challenges, partially due to the high demand of training data, specifically in terms of scale, diversity, and the precision required for alignment.

### 3 Proposed Method: LightAvatar

#### 3.1 Prerequisites: NeRF-based Avatars and NeLF

A head avatar typically has two major components during inference: *a parametric head model* and *renderer*. The parametric model describes the head geometry under different expressions. The renderer is responsible to synthesize the final head images, using the supervision of RGB data from a monocular video.

In this paper, FLAME [42] is adopted as the parametric model due to its extensive usage (notably, our method can be seamlessly generalized to other parametric models). For rendering, many top-performing implicit head avatars (*e.g.*, NerFACE [20], MonoAvatar [6]) employ a NeRF-based representation that maps a 5D input (point position and viewing direction) to a 4D output (RGB and density). In inference, multiple points along a ray are sampled. For each point, the NeRF network is queried to obtain its RGB and density. Finally, the color of each ray is obtained via alpha compositing [34, 47, 50].

One major problem preventing fast inference in NeRF and also NeRF-based avatars is that the the number of sampled points is pretty large. Consequently, the rendering computation for even a single pixel is prohibitively heavy, limiting the NeRF-based avatars usage in real-time applications.

**Neural Light Field (NeLF).** This paper attempts at solving the aforementioned slow rendering problem by introducing a *light field based* avatar. Unlike NeRF, NeLF learns a mapping from a camera ray to 3D RGB directly,  $\mathcal{F}_{\Theta} : \mathbb{R}^4 \mapsto \mathbb{R}^3$ , without any alpha compositing. Rendering with NeLFs amounts to a single network forward *vs.* hundreds of network forward passes with NeRFs, thus being much faster. Although existing works [14, 72] showed compelling NeLFs for *static* scenes, to our best knowledge, there are no successfully attempts that have utilized it for building photo-realistic head avatars (which are *dynamic*). This work is meant to bridge this gap.

### 3.2 LightAvatar: A Dynamic NeLF-based Avatar

As illustrated in Fig. 2, our LightAvatar has two inputs, *a ray representation* and *an expression representation*, and directly predicts the RGB of the ray, without explicitly relying on any geometry. The rendering process is equivalent to a single network forward. The ray representation encodes ray position information; expression representation encodes *view-dependent* expression information. In the following, we detail our design choices for both these representations.

**(1) Ray Representation.** Following prior NeLF works [14, 72], we use the *PointConcat* scheme to obtain the ray representation. Specifically, given a camera ray (its origin  $\mathbf{r}_o$  and direction  $\mathbf{r}_d$ ), we sample  $K$  evenly-spaced points along the ray between near plane and far plane. Then, these point coordinates are concatenated as a long vector,

$$\mathbf{I}_{\text{ray}} = (x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_K, y_K, z_K). \quad (1)$$

Following prior works [14, 50, 72], the coordinates in the ray representation are further transformed by positional encoding [71] to enrich their expressive power.

**(2) Expression Representation.** Differently from the rays, that capture a specific local region, the expression code itself is a global descriptor. Therefore, for different rays, the corresponding expression representation should be discriminative, depending on the specific viewing direction. Based on these considerations, the expression representation is designed to be made up of two parts, *a local feature bank* and *spatial attention weights*.

The local feature bank stores local features  $\mathbf{Z} \in \mathbb{R}^{N_{\text{lf}} \times D_{\text{lf}}}$  of different head regions (such as eyes, nose, mouth, *etc.*). It is obtained via an MLP called *local feature network* (see Fig. 2) from the expression code.

The spatial attention weights are a vector  $\mathbf{W} \in \mathbb{R}^{1 \times N_{\text{lf}}}$ , which is the output of an MLP network called *spatial attention network* (see Fig. 2). Given a specific ray and an expression code, the spatial attention network generates weights that indicate the degree of attention to be allocated to distinct spatial regions. For instance, if the ray intersects an eye region, the spatial attention network should prioritize local features of that region. This approach makes the expression representation more view-dependent. Given the local feature bank  $\mathbf{Z}$  and spatial attention weights  $\mathbf{W}$ , the final expression representation for a ray is obtained via

the following *matrix multiplication*,

$$\mathbf{I}_{\text{exp}} = \mathbf{W} \cdot \mathbf{Z}. \quad (2)$$

**NeLF Model Input.** The combination of the ray representation (Eq. (1)) and the view-dependent expression representation (Eq. (2)) constitutes the input of the NeLF model  $\mathcal{F}_{\Theta}$ . Specifically, they are concatenated together to form a single vector as the model input:

$$\mathbf{I} = \text{Concat}(\mathbf{I}_{\text{ray}}, \mathbf{I}_{\text{exp}}). \quad (3)$$

Then, the RGB is predicted via a single network forward,

$$\hat{\mathbf{c}} = \mathcal{F}_{\Theta}(\mathbf{I}). \quad (4)$$

**SR Model for Upsampling.** To reduce the computation cost, we introduce an image super-resolution (SR) network (with  $\times 4$  scale), following the NeLF network (as shown in Fig. 2):

$$I_{SR} = \mathcal{S}\mathcal{R}_{\Phi}(I_{LR}), \quad (5)$$

where  $I_{LR} \in \mathbb{R}^{h \times w \times 3}$ ,  $I_{SR} \in \mathbb{R}^{h*4 \times w*4 \times 3}$ . The  $I_{SR}$  is the final output RGB.

**Network Architecture.** The architectures of the NeLF backbone, spatial attention network, local feature network, and SR network are detailed as follows.

(i) *NeLF backbone*. It has three parts: head, body, and tail. The head is a one-layer MLP which maps the input to the internal 128-D feature. The body consists of many residual MLP blocks of width 128 (each block has two layers). The tail is also a one-layer MLP which maps the feature to RGB. There is also a skip connection between the head and tail.

(ii) *Spatial attention network and local feature network*. Similar to the NeLF model, the spatial attention network and local feature network also have three parts: head, body, and tail. The difference is that the body is much shallower, made by only 2 residual blocks with no skip connection used between the head and tail. The final output activation of spatial attention network is *Sigmoid* to ensure that the outputs are in the range (0, 1).

The local feature bank of the local feature network is designed as a matrix of  $\mathbb{R}^{64 \times 128}$ , i.e.,  $N_{lf} = 64$ ,  $D_{lf} = 128$ . Ideally, a large local feature bank leads to more fine-grained local features. However, in practice, we do not observe significant performance improvement when using an excessively large local feature bank, while causing the cost of increased peak memory and inference time.

(iii) *SR network architecture*. Similar to the NeLF backbone, the SR network also consists of three parts: head, body, and tail. The head is a single Conv layer. The body consists of multiple residual blocks (10 in this work). In the tail, there are two upsampling layers for super-resolution and another Conv layer to project back to RGB. We use *Transpose Conv* layers (stride = 2) for  $\times 2$  upsampling, inspired by prior work MobileR2L [14]. Yet importantly, MobileR2L [14] distributes the upsampling layers throughout the body of the SR network. This

would lead to sizable increase of FLOPs because after each upsampling, the feature map size doubles, meaning the FLOPs scales by a factor of  $4\times$  thereafter. Instead, we defer the upsampling to the lightweight tail, tapping into the modern SR architecture design wisdom [43]. By doing so, the major backbone of the SR network maintains a compact spatial feature map size, which is critical to the ultra-low FLOPs of our model (Tab. 3).

**Incorporating Shoulders in the Head Avatar.** Head and shoulders can rotate in different directions. Such a morphable structure creates extra challenges for digitizing a human. Prior related works typically tackle the head and shoulder or torso separately, *i.e.*, using two components to render the head and shoulders respectively, while we aim at modeling the two parts with a single model.

Specifically, to capture and render the shoulders, we simply add a new set of rays representation *using the shoulder rotation* tracked by off-the-shelf fitting algorithms [42]. This simply means the NeLF network input in Eq. (1) becomes a *longer* vector, no more extra design needed. Essentially, this design implies that we believe the model has sufficient capacity and the input we provide has sufficient information to predict the RGB of both head and shoulders.

### 3.3 Training via Distillation

Despite the attractive simplicity of our method, our preliminary results suggest that training the proposed LightAvatar model on the original data does not offer satisfactory performance. This problem can be solved by training the model with *sufficient* data. Thus, we propose to employ a *pretrained* teacher model (MonoAvatar [6]) to synthesize more pseudo data. Specifically, given a pseudo input (expression  $\mathbf{e}$  and camera pose - camera origin  $\mathbf{r}_o$  and rotation  $\mathbf{R}$ ), we query the teacher model  $\mathcal{T}$  to obtain the output RGB image,

$$\mathbf{c}^{(t)} = \mathcal{T}(\mathbf{e}, \mathbf{r}_o, \mathbf{R}), \quad (6)$$

where the expression and camera pose are obtained by *interpolating* two frames randomly drawn from the real data:

$$\begin{aligned} \mathbf{e} &= \alpha \cdot \mathbf{e}_1 + (1 - \alpha) \cdot \mathbf{e}_2, \\ \mathbf{r}_o &= \alpha \cdot \mathbf{r}_{o1} + (1 - \alpha) \cdot \mathbf{r}_{o2}, \\ \mathbf{R} &= \alpha \cdot \mathbf{R}_1 + (1 - \alpha) \cdot \mathbf{R}_2, \end{aligned} \quad (7)$$

where  $\alpha \sim U[0, 1]$ . The training example of pseudo data is organized as,

$$(\mathbf{r}_o, \mathbf{r}_d, \mathbf{e}, \mathbf{c}^{(t)}), \quad (8)$$

where  $\mathbf{r}_o, \mathbf{r}_d, \mathbf{e}$  represent the ray origin, direction, and expression, respectively. These together make the the input of our LightAvatar model. For the real data, we use a similar format and replace  $\mathbf{c}^{(t)}$  with the real captured RGB  $\mathbf{c}$ .

**Loss function.** Our LightAvatar model  $\mathcal{F}_{\Theta}$  is trained with the photometric loss that minimizes the  $L_2$  distance between the predicted image  $I_{SR}$  and target

image  $I_{target}$  and a perceptual loss [33],

$$\mathcal{L} = \|I_{SR} - I_{target}\|_2^2 + \lambda \mathcal{P}(I_{SR}, I_{target}), \quad (9)$$

where  $I_{target}$  can come from pseudo image or real image;  $\mathcal{P}$  means perceptual loss with VGG network [62]. The coefficient  $\lambda$  is the loss weight (0.005 by default).

### 3.4 Warping Field Network

This section introduces a warping field network, inspired by prior works [6, 32, 75], to correct the fitting noise in the real data. The warping field network takes a *trainable* per-frame latent variable (denoted as  $\mathbf{v}_i$ ,  $i \in [N]$ ,  $N$  is the number of training frames) and point position ( $\mathbf{q}$ ) as input, and output a new point position (*i.e.*, the warped point position, denoted as  $\mathbf{q}'$ ) by the following formulation [6],

$$\begin{aligned} \mathbf{R}, \mathbf{c}^{(\text{rot})}, \mathbf{t} &= \mathcal{G}(\mathbf{q}, \mathbf{v}_i), \\ \mathbf{q}' &= \mathbf{R}(\mathbf{q} + \mathbf{c}^{(\text{rot})}) - \mathbf{c}^{(\text{rot})} + \mathbf{t}, \end{aligned} \quad (10)$$

where  $\mathbf{R}, \mathbf{c}^{(\text{rot})}, \mathbf{t}$  stands for rotation matrix, rotation center, translation, respectively;  $\mathcal{G}$  is the warping network, which is designed as a residual MLP, similar to the LightAvatar backbone *but much shallower* (we empirically find a deep warping field network is hard to converge).

During training, we mix the pseudo and real images for the best performance. Notably, the warping field network is only used for the *real* images. For pseudo images, since they are exactly aligned, they will not go through the warping field network. During testing, the warping field network is *not* needed, either.

## 4 Experimental Results

**Implementation Details.** We use TensorFlow [1] for the major experiments. Adam optimizer [39] is used with an exponential decay learning rate (LR) schedule (initial LR 5e-4, decayed by a multiplier 0.2 every 500K iterations). During training, we train the model without SR network first with ray-based pseudo data (16,384 rays per batch) for around 500K iterations. This provides the initial weights. Next, we add the SR network to jointly optimize for another 500K iterations with initial LR 1e-4, where the training data is image-based (16 images per batch). When finetuning with the real data (mixed with pseudo data), the initial LR is set to even smaller (1e-5) to avoid over-optimization.

Our NeLF model has 10 residual MLP blocks (width 128 neurons). Each residual MLP block has 2 MLP layers. The SR model has 5 residual Conv blocks (width 56 filters, kernel size  $3 \times 3$ , padding 1) with two  $\times 2$  upsample layers (thus, the total upsample scale is  $\times 4$ ). The total FLOPs of the whole model is designed to be ultra low: **0.09M per pixel**, which is an order-of-magnitude smaller than existing counterparts (see Tab. 3). Code is released at: <https://github.com/MingSun-Tse/LightAvatar-TensorFlow>.

**Table 1:** LPIPS $\downarrow$ /SSIM $\uparrow$ /PSNR $\uparrow$  comparison with prior qualitatively top-performing methods. The **best** results are in **bold**, second best underlined.

Method	<i>Subject 0</i>		<i>Subject 1</i>		<i>Subject 2</i>		<i>Subject 3</i>		<i>Subject 4</i>		<i>Average</i>
	LPIPS	SSIM									
TPSMM	0.192/0.852/22.60		0.205/0.830/16.38		0.216/0.782/18.40		0.222/0.799/20.28		0.156/0.913/21.29		0.198/0.835/19.79
FOMM	0.171/0.841/22.93		0.179/0.827/16.02		0.202/0.777/18.98		0.186/0.798/22.28		0.122/0.915/23.94		0.172/0.832/20.83
NHA	0.165/0.836/20.20		0.166/0.840/15.48		0.178/0.809/17.99		0.153/0.798/21.31		0.091/0.926/23.78		0.151/0.842/19.75
IMAvatar	0.207/0.852/21.26		0.187/0.848/15.98		0.265/0.729/15.80		0.214/0.782/20.37		0.142/0.897/20.63		0.203/0.822/18.81
NerFACE	0.205/0.817/20.06		0.182/0.833/15.78		0.188/0.793/19.41		0.229/0.747/18.16		0.093/0.938/25.57		0.179/0.826/19.80
MonoAvatar	0.144/0.864/21.92		0.152/0.855/16.23		0.141/0.841/20.42		0.156/0.833/23.05		0.075/0.944/25.71		0.134/0.867/21.47
Ours	0.136/0.864/22.12		0.138/0.855/16.32		0.117/0.844/20.82		0.137/0.836/23.43		0.060/0.947/25.88		<b>0.118/0.869/21.71</b>

**Datasets.** We compare our approach to others on **14** monocular videos of different subjects, named *Subject0* to *Subject13* in this work. *Subject0* to *Subject12* are from prior works like NerFACE [20] and MonoAvatar [6]. *Subject13* is captured by this work, with shoulder – we shall show our method can reliably model the shoulder on this subject. The backgrounds of these videos are removed and only heads (and shoulder for *Subject13*) are retained. We resize the video to make sure the longer side is 512. Each video is split into two parts. The first part is used for training and the other part reserved for testing. We choose MonoAvatar [6] as teacher to synthesize 20K pseudo frames (Sec. 3.3).

**Comparison Methods and Evaluation Metrics.** We compare with existing popular head avatar methods: FOMM [61], TPSMM [82], NHA [26], IMAvatar [83], NerFACE [20], and MonoAvatar [6]. Among them, MonoAvatar is the prior SOTA in terms of quality (thus chosen as our teacher). Besides, some very recent works also focus on fast avatars as we do, *e.g.*, NeRFBlendShape [21], PointAvatar [84], and INSTA [85]. Our work is fundamentally different from them in that we build upon a different representation (neural light fields instead of NeRFs or point clouds). We shall also compare with them.

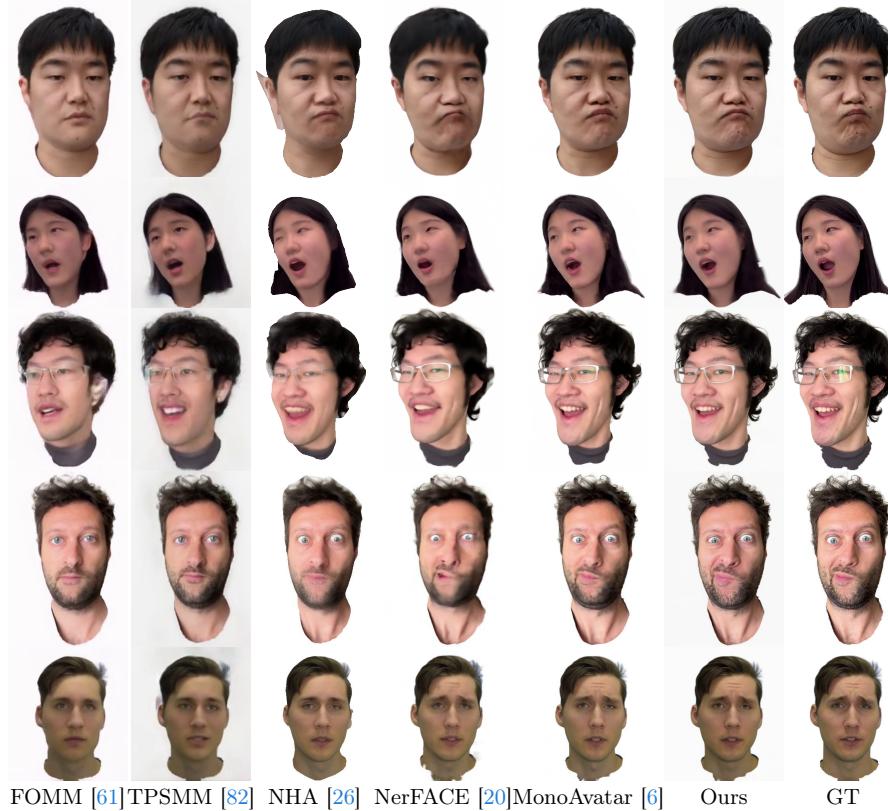
We evaluate different methods with standard metrics: LPIPS [80] / SSIM [73] / PSNR. Of note, it is well-known that LPIPS capture the structural details more accurately than the pixel-wise PSNR and patch-based SSIM. Thus, when evaluating the quantitative results, it is advised to put more weight on LPIPS.

#### 4.1 Comparison with Other Approaches

**Quantitative Comparison.** The quantitative comparisons are presented in Tab. 1 and Tab. 2. Our LightAvatar consistently achieves the *best* average LPIPS/SSIM/PSNR in two tables, showing the encouraging potential of employing light fields to represent head avatars.

Notably, although we use MonoAvatar [6] as teacher to synthesize pseudo data, our final results actually *surpass* the teacher (see Tab. 1). This phenomenon agrees with the previous observations in the static neural light fields [14, 72]. This is because the pseudo data (*i.e.*, the supervision from the teacher) only provides the initial weights to our model for the subsequent finetuning on the real data. The performance of our model is *not* bounded by the teacher.

**Qualitative Comparison.** The qualitative results are shown in Fig. 3 and Fig. 4. (1) As seen, although our method does not rely on explicit mesh, it faithfully predicts the nuanced facial expressions like those that explicitly use



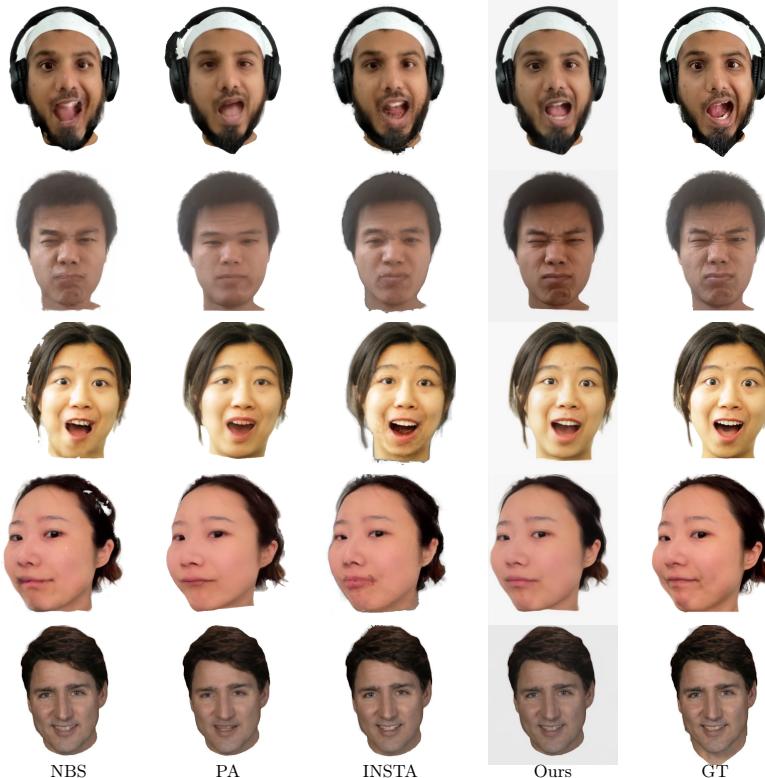
**Fig. 3:** Visual comparison on the test set with prior top-performing monocular head avatars. From top to down, the subject is *Subject0* to *Subject4* in order (see another 3 in supplementary material). Our LightAvatar method faithfully predicts the facial expressions and presents *sharper* high-frequency details than other approaches.

mesh (such as MonoAvatar [6]). LightAvatar is slightly better than MonoAvatar in general, as it produces less grain noise (*e.g.*, zoom in and compare the eyes and mouth region of *Subject 1*). (2) Compared to the fast avatars, our LightAvatar generates obviously better quality. According to the results, PointAvatar [84] often distorts the facial expressions. INSTA [85] produces quite many artifacts around the mouth area. NeRFBlendShape [21] sometimes produce incomplete structures (like the hair in *Subject 10* and *Subject 11*).

**Inference and Training Efficiency.** (1) Tab. 3 shows the model complexity (FLOPs and run-time speed) comparison. Our method is *significantly faster* than the teacher MonoAvatar [6] and other fast avatars methods, thanks to our dedicated light field design mixed with 2D SR ConvNet. Notably, even INSTA [85] employs customized C++/CUDA implementation, which is usually faster than pure TensorFlow or PyTorch implementation, our method is still significantly faster than INSTA, owing to the ultra-low FLOPs.

**Table 2:** LPIPS $\downarrow$ /SSIM $\uparrow$ /PSNR $\uparrow$  comparison with recent *fast avatar* methods. NBS: NeRFBlendShape [21], PA: PointAvatar [84], INSTA [85].

Method	Subject 8			Subject 9			Subject 10			Subject 11			Subject 12			Average
	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	
NBS	0.093/0.882/20.50	0.104/0.933/25.99	0.108/0.900/25.66	0.081/0.924/24.78	0.049/0.960/23.78	0.087/0.920/24.14										
PA	0.109/0.839/19.51	0.119/0.913/24.20	0.132/0.853/22.51	0.104/0.903/23.11	0.068/0.939/22.98	0.106/0.889/22.46										
INSTA	0.124/0.857/20.65	0.142/0.910/23.3	0.173/0.833/20.39	0.102/0.913/24.28	0.058/0.951/23.93	0.120/0.893/22.51										
Ours	0.067/0.888/21.50	0.098/0.936/25.70	0.093/0.896/24.97	0.072/0.929/25.33	0.045/0.960/23.94	0.075/0.922/24.29										



**Fig. 4:** Visual comparison with recent fast avatars. NBS: NeRFBlendShape [21], PA: PointAvatar [84], INSTA [85]. Top to down: *Subject 8* to *Subject 12*.

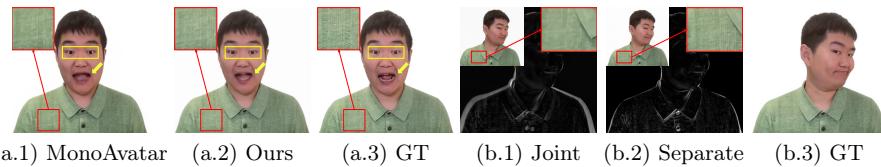
(2) This work does not focus on improving the training speed. It takes around 20 hrs with 4 V100 (16GB) GPUs to train our model of one subject with TensorFlow, which is only 1/5 cost of previous NeLF papers on static scenes [72], thanks to our ultra lightweight model design. For reference, PointAvatar reports  $\sim$ 6 hrs with 1 A100 (80G), comparable to our cost considering A100 is around 2  $\sim$  3 $\times$  faster than V100. INSTA markedly reports  $\sim$ 10 mins with 1 RTX 3090 GPU, yet it comes with an inferior quality and slower inference than ours.

#### 4.2 Results with Shoulders

In Fig. 5(a), we present the results with the shoulders to show the capability of our method in capturing the torso area. Our approach achieves *better* test LPIPS/SSIM/PSNR compared to the teacher MonoAvatar [6]. Visually, our

**Table 3:** FLOPs per pixel (estimated for rendering a  $512 \times 512$  image) and run-time speed (fps) comparison on an NVIDIA GeForce RTX 3090 GPU (24GB). Compared to the teacher, we achieve speedup by *two orders of magnitude*. Note, our method is not only much faster, but also superior in terms of image quality (see Tab. 2, Fig. 4).

Method	FLOPs (M)	Implementation	Speed (fps)
MonoAvatar [6] (teacher)	9.10	TensorFlow	0.5
PointAvatar [84]	3.56	PyTorch	5.0
INSTA [85]	1.01	C++/CUDA	46.2
NeRFBlendShape [21]	0.85	PyTorch	11.2
Ours	<b>0.09</b>	PyTorch	<b>174.1</b>



**Fig. 5:** (a) Results of our method on *Subject13* with the shoulder. For reference, the average LPIPS/SSIM/PSNR of Monoavatar on test set: 0.118/0.846/26.11; ours: 0.107/0.849/26.29. (b) Comparison between *joint modeling* and *separate modeling* (Sec. 3.2) when learning the shoulder in our method.

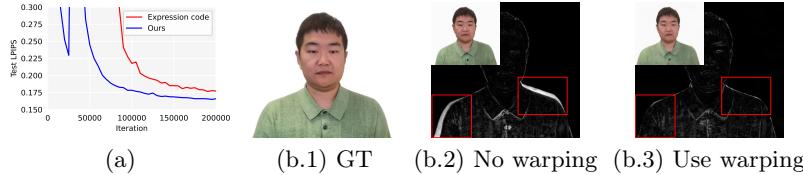
method also produces *sharper* details (such as the textures of the apparel and the reflections in the irises). Notably, for the inside-mouth area (note the yellow arrow), it is not covered by the head mesh. MonoAvatar thus produces more artifacts since it anchors local radiance fields to mesh vertices. In contrast, our mouth area is more visually pleasing since we do not rely on mesh.

In Fig. 5(b), we show the effect of using the proposed *separate modeling* scheme (Sec. 3.2) to learn the shoulder area in our LightAvatar. As seen, if the torso rotation is not considered, the shoulder will mistakenly rotate with the head rotation, causing severe miss-alignments and flickering issues. Instead, with the proposed separate modeling scheme, LightAvatar learns to disentangle the neck rotation from the rest of the body. As such, the rendered shoulder is not flickering anymore. This shows the encouraging potential of our method to handle different parts of an avatar by a single model.

### 4.3 Ablation Study

**Effect of our Expression Representation.** A naive baseline for the expression representation is to use the raw expression code (with positional encoding). Fig. 6(a) presents the comparison between our proposed expression representation and this naive scheme. As seen, our expression representation helps the LightAvatar model converge much faster and achieve a better test LPIPS.

**Effect of Warping Field Network.** Fig. 6(b) shows the comparison of using and not using the warping field network. The warping field network makes the rendered image more aligned with the ground-truth. With the warping net-



**Fig. 6:** (a) Test LPIPS comparison between using the proposed expression representation (Eq. (2)) and the raw expression code in our LightAvatar on *Subject 0*. The two models are trained for the same iterations ( $200K$ ). (b) Comparison between *not using* and *using* the proposed *warping field network* (Sec. 3.4) in our method. The black-background image refers to the pixel-wise difference between the predicted image and the ground-truth, where *brighter* color indicates *larger* difference.

work, we can finetune the model on real data, gaining better quality while not undermining the temporal consistency.

## 5 Conclusion and Limitations

This work introduces *LightAvatar*, offering a compelling proof-of-concept for building photo-realistic head avatars with neural *light* fields. LightAvatar features a simple and uniform network design - it takes 3DMM parameters and camera pose as input and predicts the image via a single network forward pass. We introduce dedicated network designs to ensure training stability and high rendering efficiency. The model training is challenging due to the light field formulation and compact network design. To resolve this, we present a *distillation-based* training pipeline and a warp field network so as to mitigate the fitting error in the real data. Extensive results and analyses on many subjects show our LightAvatar reaches SOTA image quality while being significantly faster than the counterparts, rendering at **174.1 FPS** on a consumer-grade GPU (RTX3090).

*Limitations.* This work still has several limitations to overcome. (1) We do not explicitly consider the more nuanced expressions (such as the eye-ball rotation) in our current framework. The model could learn these variations by itself, but we still observe a few small differences *vs.* the ground-truth (Fig. 9(a) in supp.). Besides, it is also challenging for our method to handle complex structures like long hairs (Fig. 9(b) in supp.), similar to existing top-performing NeRF-based avatars. How to integrate these nuanced details in our framework to make the avatar even more photo-realistic is a worthy next step. (2) This work aimed at building a NeLF-based avatar for faster *inference* in this paper. The *training* efficiency of building the avatars is not improved much in this work (yet still comparable to counterparts such as PointAvatar [84]). Accelerating training (like INSTA [85]) is also worth exploring in the future. (3) 3DGS-based avatars (*e.g.*, GaussianAvatars [56]) are competitive approaches for fast rendering. Although our method is much faster than GaussianAvatars [56] ( Tab. 5 in supp.), a more comprehensive comparison in terms of the quality is preferred in the future.

## References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: TensorFlow: A system for large-scale machine learning. In: IEEE Symposium on Operating Systems Design and Implementation (2016) [9](#)
2. Athar, S., Xu, Z., Sunkavalli, K., Shechtman, E., Shu, Z.: Rignerf: Fully controllable neural 3d portraits. In: CVPR 2022. pp. 20332–20341 (2022) [2](#), [3](#)
3. Attal, B., Huang, J.B., Zollhöfer, M., Kopf, J., Kim, C.: Learning neural light fields with ray-space embedding. In: CVPR (2022) [5](#)
4. Bai, Z., Cui, Z., Liu, X., Tan, P.: Riggable 3d face reconstruction via in-network optimization. In: CVPR 2021. pp. 6216–6225 (2021) [3](#)
5. Bai, Z., Cui, Z., Rahim, J.A., Liu, X., Tan, P.: Deep facial non-rigid multi-view stereo. In: CVPR 2020. pp. 5849–5859 (2020) [3](#)
6. Bai, Z., Tan, F., Huang, Z., Sarkar, K., Tang, D., Qiu, D., Meka, A., Du, R., Dou, M., Orts-Escalano, S., et al.: Learning personalized high quality volumetric head avatars from monocular rgb videos. In: CVPR (2023) [2](#), [3](#), [5](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [20](#), [21](#), [22](#)
7. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021) [3](#), [4](#)
8. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR (2022) [3](#), [4](#)
9. Bemana, M., Myszkowski, K., Seidel, H.P., Ritschel, T.: X-fields: Implicit neural view-, light-and time-image interpolation. ACM Transactions on Graphics **39**(6), 1–15 (2020) [5](#)
10. Bharadwaj, S., Zheng, Y., Hilliges, O., Black, M.J., Fernandez-Abrevaya, V.: Flare: Fast learning of animatable and relightable mesh avatars. In: SIGGRAPH Asia (2023) [4](#)
11. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999) [2](#), [3](#)
12. Buciluă, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: SIGKDD (2006) [2](#)
13. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Trans. Graph. **35**(4), 126:1–126:12 (2016) [3](#)
14. Cao, J., Wang, H., Chemerys, P., Shakhrai, V., Hu, J., Fu, Y., Makoviichuk, D., Tulyakov, S., Ren, J.: Real-time neural light field on mobile devices. In: CVPR (2023) [2](#), [5](#), [6](#), [7](#), [10](#), [20](#), [24](#)
15. Chaudhuri, B., Vesdapunt, N., Shapiro, L.G., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: ECCV 2020. vol. 12350, pp. 142–160 (2020) [3](#)
16. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: ECCV 2022. vol. 13692, pp. 333–350 (2022) [5](#)
17. Chen, Z., Funkhouser, T.A., Hedman, P., Tagliasacchi, A.: Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In: CVPR 2023. pp. 16569–16578. IEEE (2023) [2](#), [5](#)
18. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: CVPR (2019) [2](#)
19. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Trans. Graphic. **39**(5), 1–38 (2020) [2](#), [3](#)

20. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR (2021) [2](#), [3](#), [5](#), [10](#), [11](#)
21. Gao, X., Zhong, C., Xiang, J., Hong, Y., Guo, Y., Zhang, J.: Reconstructing personalized semantic facial nerf models from monocular video. ACM Trans. Graph. **41**(6), 200:1–200:12 (2022) [4](#), [10](#), [11](#), [12](#), [13](#)
22. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: High-fidelity neural rendering at 200fps. arXiv preprint arXiv:2103.10380 (2021) [2](#), [5](#)
23. Garrido, P., Valgaerts, L., Rehmsen, O., Thormählen, T., Pérez, P., Theobalt, C.: Automatic face reenactment. In: CVPR 2014. pp. 4217–4224 (2014) [3](#)
24. Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., Theobalt, C.: Reconstruction of personalized 3d face rigs from monocular video. ACM Trans. Graph. **35**(3), 28:1–28:15 (2016) [3](#)
25. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (1996) [5](#)
26. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: CVPR (2022) [2](#), [10](#), [11](#)
27. Gupta, A., Cao, J., Wang, C., Hu, J., Tulyakov, S., Ren, J., Jeni, L.: Lightspeed: light and fast neural light fields on mobile devices. In: NeurIPS (2023) [5](#)
28. Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P.: Baking neural radiance fields for real-time view synthesis. In: ICCV (2021) [5](#)
29. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS Workshop (2014) [2](#)
30. Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y., Li, H.: Avatar digitization from a single image for real-time rendering. ACM Trans. Graph. **36**(6), 195:1–195:14 (2017) [3](#)
31. Ichim, A.E., Bouaziz, S., Pauly, M.: Dynamic 3d avatar creation from hand-held video input. ACM Trans. Graph. **34**(4), 45:1–45:14 (2015) [3](#)
32. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: ECCV (2022) [9](#)
33. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016) [9](#)
34. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. SIGGRAPH **18**(3), 165–174 (1984) [5](#)
35. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics **35**(6), 1–10 (2016) [5](#)
36. Karis, B., Games, E.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice **4**(3), 1 (2013) [4](#)
37. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM ToG **42**(4), 1–14 (2023) [23](#)
38. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Trans. Graph. **37**(4), 163 (2018) [3](#)
39. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) [9](#)
40. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ToG **39**(6), 1–14 (2020) [4](#)
41. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (1996) [5](#)
42. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. In: SIGGRAPH Asia (2017) [2](#), [5](#), [8](#)

43. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPR Workshop (2017) [2](#), [8](#)
44. Lindell, D.B., Martel, J.N., Wetzstein, G.: Autoint: Automatic integration for fast neural volume rendering. In: CVPR (2021) [5](#)
45. Liu, C., Li, Z., Yuan, J., Xu, Y.: Neulf: Efficient novel view synthesis with neural 4d light field. In: EGSR (2022) [5](#)
46. Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De La Torre, F., Sheikh, Y.: Pixel codec avatars. In: CVPR (2021) [4](#)
47. Max, N.: Optical models for direct volume rendering. TVCG **1**(2), 99–108 (1995) [5](#)
48. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR (2019) [2](#)
49. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics **38**(4), 1–14 (2019) [5](#)
50. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) [2](#), [3](#), [4](#), [5](#), [6](#)
51. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (2022) [4](#), [5](#)
52. Neff, T., Stadlbauer, P., Parger, M., Kurz, A., Mueller, J.H., Chaitanya, C.R.A., Kaplanyan, A.S., Steinberger, M.: DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks. Computer Graphics Forum (2021) [5](#)
53. Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., Huynh-The, T., Navabandi, S., Nguyen, T.T., Pham, Q.V., Nguyen, C.M.: Deep learning for deepfakes creation and detection: A survey. CVIU **223**, 103525 (2022) [24](#)
54. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR (2019) [2](#)
55. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019) [23](#)
56. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: CVPR (2024) [14](#), [23](#)
57. Rebain, D., Jiang, W., Yazdani, S., Li, K., Yi, K.M., Tagliasacchi, A.: Derf: Decomposed radiance fields. In: CVPR (2021) [2](#), [5](#)
58. Reiser, C., Peng, S., Liao, Y., Geiger, A.: Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In: ICCV (2021) [2](#), [5](#)
59. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) [24](#)
60. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016) [24](#)
61. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NeurIPS (2019) [10](#), [11](#)
62. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) [9](#)

63. Sitzmann, V., Rezchikov, S., Freeman, W.T., Tenenbaum, J.B., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. In: NeurIPS (2021) 5
64. Suhail, M., Esteves, C., Sigal, L., Makadia, A.: Light field neural rendering. In: CVPR (2022) 5
65. Sun, C., Sun, M., Chen, H.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: CVPR 2022. pp. 5449–5459 (2022) 5
66. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H., Pérez, P., Zollhöfer, M., Theobalt, C.: FML: face model learning from videos. In: CVPR 2019. pp. 10812–10822 (2019) 3
67. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P.P., Tretschk, E., Wang, Y., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Nießner, M., Barron, J.T., Wetzelstein, G., Zollhöfer, M., Golyanik, V.: Advances in neural rendering. Comput. Graph. Forum **41**(2), 703–735 (2022) 4
68. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: ECCV 2020. vol. 12361, pp. 716–731 (2020) 3
69. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 183:1–183:14 (2015) 3
70. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of RGB videos. In: CVPR 2016. pp. 2387–2395 (2016) 3
71. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 6
72. Wang, H., Ren, J., Huang, Z., Olszewski, K., Chai, M., Fu, Y., Tulyakov, S.: R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In: ECCV (2022) 2, 5, 6, 10, 12, 24
73. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004) 10
74. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Trans. Graph. **30**(4), 77 (2011) 3
75. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: CVPR (2022) 9
76. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: A large-scale high quality 3d face dataset and detailed rippable 3d face prediction. In: CVPR 2020. pp. 598–607 (2020) 3
77. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q.: Deep learning for single image super-resolution: A brief review. TMM **21**(12), 3106–3121 (2019) 2
78. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: ICCV (2021) 2, 5
79. Yu, H., Julin, J., Milacski, Z.A., Niiuma, K., Jeni, L.A.: Dylin: Making light field networks dynamic. In: CVPR (2023) 2
80. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 10
81. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018) 2
82. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR (2022) 10, 11

83. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: Im avatar: Implicit morphable head avatars from videos. In: CVPR (2022) [2](#), [3](#), [10](#)
84. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: CVPR 2023. pp. 21057–21067. IEEE (2023) [4](#), [10](#), [11](#), [12](#), [13](#), [14](#)
85. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: CVPR 2023. pp. 4574–4584 (2023) [2](#), [4](#), [10](#), [11](#), [12](#), [13](#), [14](#)
86. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. Comput. Graph. Forum **37**(2), 523–550 (2018) [3](#)

## 6 Overview of Supplementary Materials

In the following supplementary materials, we provide:

- more detailed network architecture explanations (Sec. 7);
- comparisons with the teacher on three more subjects (Sec. 8.1);
- video comparisons with different methods (Sec. 8.3);
- speed comparison with a Gaussian-Splatting-based avatar (Sec. 8.4);
- video comparison between the joint modeling scheme and separate modeling when learning the shoulder (Sec. 8.5);
- ablation study: visual comparison between using our expression representation *vs.* using the raw expression code as NeLF input (Sec. 8.6);
- 3D consistency check of our method (Sec. 8.7);
- potential negative impact discussion (Sec. 8.8).

## 7 Detailed Network Architectures

Two types of networks are used in our method, MLP network and Conv network. Their architectures, as shown in Fig. 7, follow a pretty simple and uniform paradigm, by our particular design.

For the SR model, we use the Conv network architecture; for the others, we use the MLP network architecture. Major differences lie in the number of residual blocks (ResBlocks) used in the body (which determines the network depth), and the width for internal layers. For NeLF network, we use 10 ResBlocks with width 128. For spatial attention network and local feature network, we use 2 ResBlocks with width 128. For SR network, we use 5 ResBlocks with width 56.

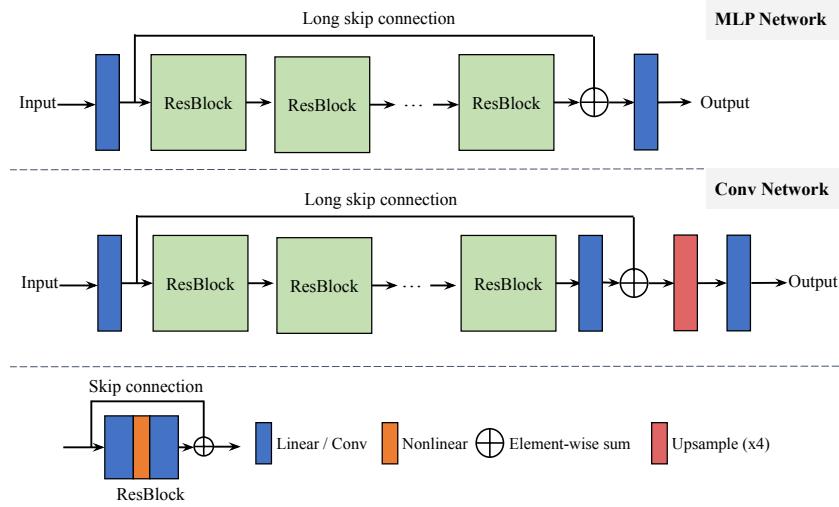
LeakyReLU is used for the major non-linearity (unless specified otherwise). We do *not* use any normalization layers (unlike prior related works like MobileR2L [14]) because (1) our specially designed training strategy already makes the algorithm converge successfully and fast; (2) in practice, normalization layers will lead to extra inference latency.

## 8 Additional Results

### 8.1 Results on *Subject 5* to *Subject 7*

The quantitative results of the other three subjects (*Subject 5* to *Subject 7*) are presented in Tab. 4, visual results in Fig. 8. Here we mainly compare with MonoAvatar [6] (the teacher), since it has been shown to be better than the other methods in Tab. 1.

As seen, similar to the results (Tab. 1, Fig. 3) in the main paper, our LightAvatar still consistently achieves *better* quantitative results and slightly better (or comparable) visual quality than the teacher. Here we also observe that our LightAvatar produces less grain noise than MonoAvatar, *e.g.*, the mouth area of *Subject 6*. Note, we achieve these better or comparable results with more than 300 $\times$  speedup against MonoAvatar (see Tab. 3 in the paper).



**Fig. 7:** Illustration of two types of networks used in our method. We intentionally use simple and uniform network designs for the models in our method - Each model has 3 parts: head, body, and tail, with a long skip connection bypassing the body.

For *Subject 5*, both MonoAvatar and our method do not accurately reconstruct the complex long hair structure and extreme facial expression, which is considered as a limitation of our method now (as discussed in the paper, Sec. 5).

**Table 4:** LPIPS $\downarrow$ /SSIM $\uparrow$ /PSNR $\uparrow$  comparison on *Subject 5/6/7*. Here we compare our method to the teacher MonoAvatar [6]. As seen, similar to the results (Tab. 1) in the main paper, our method is consistently better than teacher. Note, we achieve so meanwhile being more than 300 $\times$  faster than MonoAvatar (Tab. 3 in the main paper).

Method	<i>Subject 5</i>			<i>Subject 6</i>			<i>Subject 7</i>			<i>Average</i>		
	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR	LPIPS	SSIM	PSNR
MonoAvatar	0.246/0.704/16.57			0.160/0.862/21.70			0.119/0.891/23.45			0.175/0.819/20.57		
Ours	0.231/0.709/16.78			0.147/0.864/21.97			0.102/0.894/23.70			<b>0.160/0.822/20.82</b>		

## 8.2 Limitations of Our Method

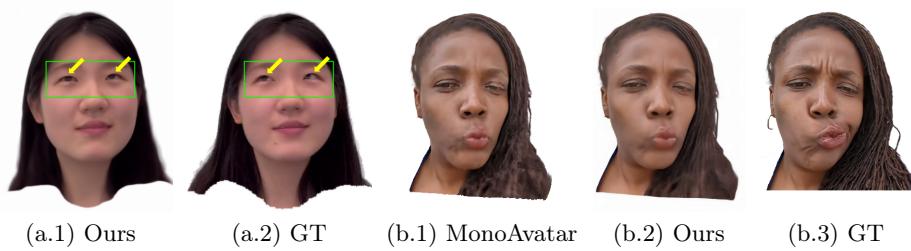
Fig. 9 showcases two limitations of our method: It is challenging to capture the nuanced details and model complex structures like long hairs.

## 8.3 Video Comparison w. Other Methods

In the paper, we presented two sets of comparisons of our method *vs.* the prior SOTA avatars (Tab. 1, Fig. 3) and the recent fast avatars specialized for fast inference speed (Tab. 2, Fig. 4). The corresponding videos are presented in the



**Fig. 8:** Visual comparison with MonoAvatar [6] on the test set. From top to down, the subject is *Subject5* to *Subject7* in order.



**Fig. 9:** Showcases of limitations of our method. (a) Our method does not explicitly integrate the eyeball rotation into the framework at present, so it is still challenging for our method to learn the correct eyeball rotation implicitly. (b) For the complex structures, such as long hairs, our method still cannot accurately reconstruct the details, similar to the NeRF-based SOTA methods (*e.g.*, MonoAvatar [6]).

attached webpage [index.html](#) (see “1. Video Comparison with 5 Subjects”, and

“2. Video Comparison with Fast Avatars”). Please find the videos in our code repository: <https://github.com/MingSun-Tse/LightAvatar-TensorFlow>.

#### 8.4 Speed Comparison w. 3DGS-Based Avatar

Recently, 3D *Gaussian Splatting* (3DGS) [37] has become a new emerging 3D representation, which has been successfully extended to building photo-realistic head avatars, such as GaussianAvatar [56].

Our work fundamentally differs from these GS-based avatars since we build upon *light fields* instead of Gaussians. GS is also known for its fast rendering speed [37] based on rasterization. GaussianAvatar inherits this advantage. The rendering speed benchmarks in Tab. 5 show that LightAvatar runs more slowly as the resolution goes higher (since the computation becomes proportionally larger *w.r.t.* output resolution), while GaussianAvatar keeps a nearly constant speed (since GA uses rasterization for rendering, which has been highly optimized on modern GPUs; the rendering speed is barely bound by the output resolution). As a result, at small resolutions, LightAvatar is faster; at larger resolutions, GaussianAvatar is faster. Simply put, LightAvatar gets the fast speed by *algorithm designs*; GaussianAvatar gets the fast speed by *hardware support* - clearly, they follow different paths. We envision one day when neural operators are as optimized as rasterization on GPUs, the potential of our method can be more fulfilled.

**Table 5:** Speed (FPS) comparison with GaussianAvatar (GA) [56] across different output resolutions, on an RTX4090 GPU (24GB) with PyTorch [55]. We use the public code<sup>†</sup> of GA for this benchmark, speed averaged by 500 frames following the guidelines of GA.

Resolution	128x128	256x256	384x384	512x512	640x640	768x768	896x896
GA	189.1	201.0	211.3	<b>217.1</b>	<b>215.8</b>	<b>212.0</b>	<b>210.8</b>
Ours	<b>408.1</b>	<b>331.4</b>	<b>295.1</b>	186.3	119.0	82.3	62.3

<sup>†</sup><https://github.com/ShenhanQian/GaussianAvatars>

#### 8.5 Video Comparison of Joint *vs.* Separate Modeling of Shoulder

The video comparison is shown in the webpage [index.html](#) (please refer to the section “3. Video comparison: Joint *vs.* separate modeling in our method”).

Clearly, the proposed separate modeling scheme achieves *much better* temporal consistency than the joint modeling, showing the encouraging potential of our method learning the head and torso with a *single* model.

#### 8.6 Visual Comparison of Different Expression Representations

One of the key contribution claimed in the paper is the designed expression representation (*Sec. 3.2, (2) Expression Representation*) *vs.* using the raw expression

code as input to regress the target RGB. In the paper, we presented the evidence in terms of the test LPIPS (Fig. 6(a)) to show the merit of our expression representation. Here we present the visual comparison in Fig. 10 to further validate our claim – Using our expression helps reconstruct clearly sharper textures.

(a) Raw expression (b) Our expression (c) GT

**Fig. 10:** Comparison between using raw expression code (a) *vs.* our expression representation (b) for the NeLF backbone in our LightAvatar (on *Subject 0*).

## 8.7 Visual Comparison of Different Expression Representations

Fig. 11 shows the reconstructed point clouds by COLMAP [59, 60] with camera poses of the test set 434 images of *Subject 0*. As seen, although our method does not have any explicit constraint for 3D consistency (unlike NeRFs), it *learns by itself* the view consistency, thanks to the massive (20K) pseudo images and sufficient capacity. This agrees with the observation in prior static-scene NeLFs [14, 72].



**Fig. 11:** 3D point cloud reconstruction with COLMAP. As seen, our method can reconstruct consistent 3D shapes despite not using any explicit 3D designs.

## 8.8 Potential Negative Impact

This work aims to build photo-realistic head avatars, which can be used in negative technology such as “Deep Fake” [53]. This has been a general potential issue for all avatar methods. To prevent misuse, we need to ensure the right access control. More advanced security measures need to be developed at the same time as avatars become more photo-realistic, such as encryption and secure storage of the avatars, using watermarks and digital signatures.

Besides, in some contexts, *e.g.*, social VR environments, the use of head avatars could raise privacy concerns if personal information or sensitive data is collected or exposed without users’ consent. To prevent this, ethical guidelines and collaborative efforts from the society are indispensable.