

# NeRFlame: FLAME-based conditioning of NeRF for 3D face rendering

Wojciech Zająć

Faculty of Mathematics and Computer Science, Jagiellonian University, Poland  
 wojciech3.zajac@student.uj.edu.pl

Jacek Tabor

Faculty of Mathematics and Computer Science, Jagiellonian University, Poland  
 jacek.tabor@uj.edu.pl

Maciej Zięba

Department of Artificial Intelligence, University of Science and Technology Wrocław, Poland  
 maciej.zieba@pwr.edu.pl

Przemysław Spurek

Faculty of Mathematics and Computer Science, Jagiellonian University, Poland  
 przemyslaw.spurek@uj.edu.pl

## Abstract

Traditional 3D face models are based on mesh representations with texture. One of the most important models is FLAME (Faces Learned with an Articulated Model and Expressions), which produces meshes of human faces that are fully controllable. Unfortunately, such models have problems with capturing geometric and appearance details. In contrast to mesh representation, the neural radiance field (NeRF) produces extremely sharp renders. But implicit methods are hard to animate and do not generalize well to unseen expressions. It is not trivial to effectively control NeRF models to obtain face manipulation.

The present paper proposes a novel approach, named NeRFlame, which combines the strengths of both NeRF and FLAME methods. Our method enables high-quality rendering capabilities of NeRF while also offering complete control over the visual appearance, similar to FLAME.

Unlike conventional NeRF-based architectures that utilize neural networks to model RGB colors and volume density, NeRFlame employs FLAME mesh as an explicit density volume. As a result, color values are non-zero only in the proximity of the FLAME mesh. This FLAME backbone is then integrated into the NeRF architecture to predict RGB colors, allowing NeRFlame to explicitly model volume density and im-

plicitly model RGB colors.

Original position



Modify positions

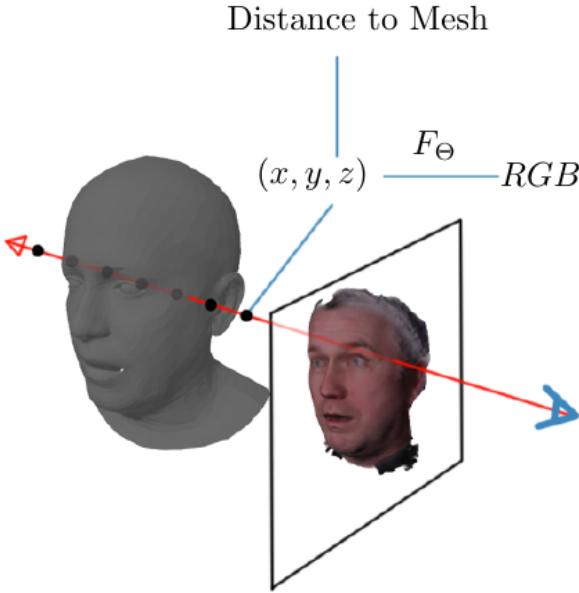


Rysunek 1. Our model allows producing manipulation of the human face. In NeRFlame, we use FLAME as a backbone, and therefore we can manipulate FLAME features and modify NeRF representation.

## 1. Introduction

Methods to automatically create fully controllable human face avatars have many applications in VR/AR, games, and telepresence. Traditional 3D face models are based on fully controllable mesh representations. FLAME [25] is one of the most important methods for

mesh-based avatars. FLAME uses a linear shape space trained from 3800 scans of human heads. FLAME combines this linear shape space with an articulated jaw, neck, eyeballs, pose-dependent corrective blend shapes, and additional global expression blend shapes. In practice, we can easily train FLAME on the 3D scan (or 2D image) of human faces and then manipulate basic behaviors like jaw, neck, and eyeballs. We can also produce colors for mesh by using textures. Unfortunately, such models have problems with capturing geometric and appearance details.



Rysunek 2. NeRFFlame architecture uses two main components: FLAME model and NeRF-based implicit representation of colors. For position  $(x, y, z)$ , we use distance to the mesh to obtain  $\sigma$  and NeRF architecture to obtain color.

Alternatively to the classical approaches, we can use implicit methods representing avatars by neural networks. NeRFs [29] represent a scene using a fully-connected architecture. As the input, NeRF takes a 5D coordinate (spatial location  $\mathbf{x} = (x, y, z)$  and viewing direction  $\mathbf{d} = (\theta, \psi)$ ) and it outputs an emitted color  $\mathbf{c} = (r, g, b)$  and volume density  $\sigma$ . NeRF extracts information from unlabeled 2D views to obtain 3D shapes. NeRF allows synthesizing of novel views of complex 3D scenes from a small subset of 2D images. Based on the relations between those base images and computer graphics principles, such as ray tracing, this neural network model can render high-quality images of 3D objects from previously unseen viewpoints. In contrast to mesh representation, NeRF captures geometric and appearance details; see Fig. 8. But it is not trivial to

effectively control NeRF to obtain face manipulation.

There exist many approaches for controlling NeRF, which use generative models [21, 9, 38], dynamic scene encoding [22], or conditioning mechanisms [4]. But still, we are not able to control NeRF as well as we can manipulate mesh representations.

This paper proposes a hybrid of NeRF and FLAME called NeRFFlame<sup>1</sup>, see Fig. 2. Our method inherits the best features from the above approaches. We can model the quality of NeRF rendering and fully control the appearance as in FLAME. Our model's fundamental idea is to explicitly model volume density without a neural network. The volume density is non-zero only in the  $\varepsilon$  neighborhood of FLAME mesh. Therefore, it is expressed by closed formula dependent on the distance to mesh. On the other hand, we use NeRF-based architecture to model RGB colors. Such a solution allows for obtaining similar quality renders to NeRF and a level of controlling mesh similar to FLAME.

Contrary to Dynamic Neural Radiance Fields, NeRFFlame is trained on one position of the human face instead of using movies with many positions. Thank FLAME backbone, we can model new unseen facial expressions. Therefore we compare our model with classical static Neural Radiance Fields.

The contributions of this paper are significant and are outlined as follows:

Firstly, we introduce NeRFFlame – an innovative hybrid of NeRF and FLAME that combines the best features of both methods, namely the exceptional rendering quality of NeRF and the precise control over appearance as in FLAME.

Secondly, we demonstrate the ability to explicitly model volume density in NeRF by employing mesh representation, which represents a significant advancement over traditional NeRF-based approaches that rely on neural networks.

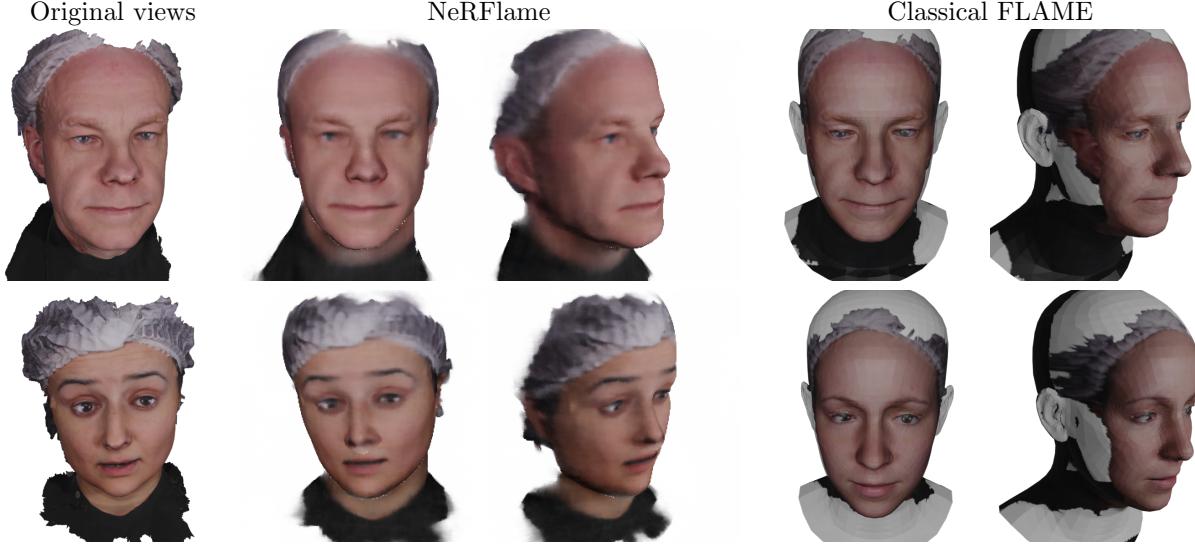
Thirdly, we train our model on a single position of the human face rather than using entire movies, thereby highlighting the versatility and practicality of our approach.

Overall, our contributions offer a significant step forward in 3D facial modeling and rendering and pave the way for future research in this area.

## 2. Related Works

NeRFFlame is a model of controllable human face avatars trained on a single 3D face represented by a few 2D images. Since we train our model on 2D images of the face, we naturally refer to Static Neural Radiance Fields. But we also can modify NeRF since we use

<sup>1</sup>The source code is available at: <https://github.com/WojtekZ4/NeRFFlame>.



Rysunek 3. Competition between NeRFlame and classical FLAME fitting. NeRF-based model better fits human expression of the face.

FLAME as a backbone. Therefore in related work, we also add Dynamic Neural Radiance Fields.

**Static Neural Radiance Fields** 3D objects can be represented by using many different approaches, including voxel grids [11], octrees [19], multi-view images [3, 27], point clouds [1, 34, 41], geometry images [35], deformable meshes [18, 25], and part-based structural graphs [24].

The above representations are discreet, which causes some problems in real-life applications. In contrast to such apprehension, NeRF [29] represents a scene using a fully-connected architecture. NeRF and many generalizations [7, 8, 26, 31, 32, 36, 39] synthesize novel views of a static scene using differentiable volumetric rendering.

One of the largest limitations is training time. To solve such problems in [15], authors propose Plenoxels, a method that uses a sparse voxel grid storing density and spherical harmonics coefficients at each node. The final color is the composition of tri-linearly interpolated values of each voxel. In [30], authors use a similar approach, but e space is divided into an independent multilevel grid. In [10], authors represent a 3D object as an orthogonal tensor component. A small MLP network, which uses orthogonal projection on tensors, obtains the final color and density. There exist some methods which use additional information to Nerf, like depth maps or point clouds [6, 13, 32, 40].

In our paper, we produce a new NeRF-based representation of 3D objects. As input, we use classical 2D images. Similarly, RGB colors are described by MLP, but volume density is given by distance to FLAME

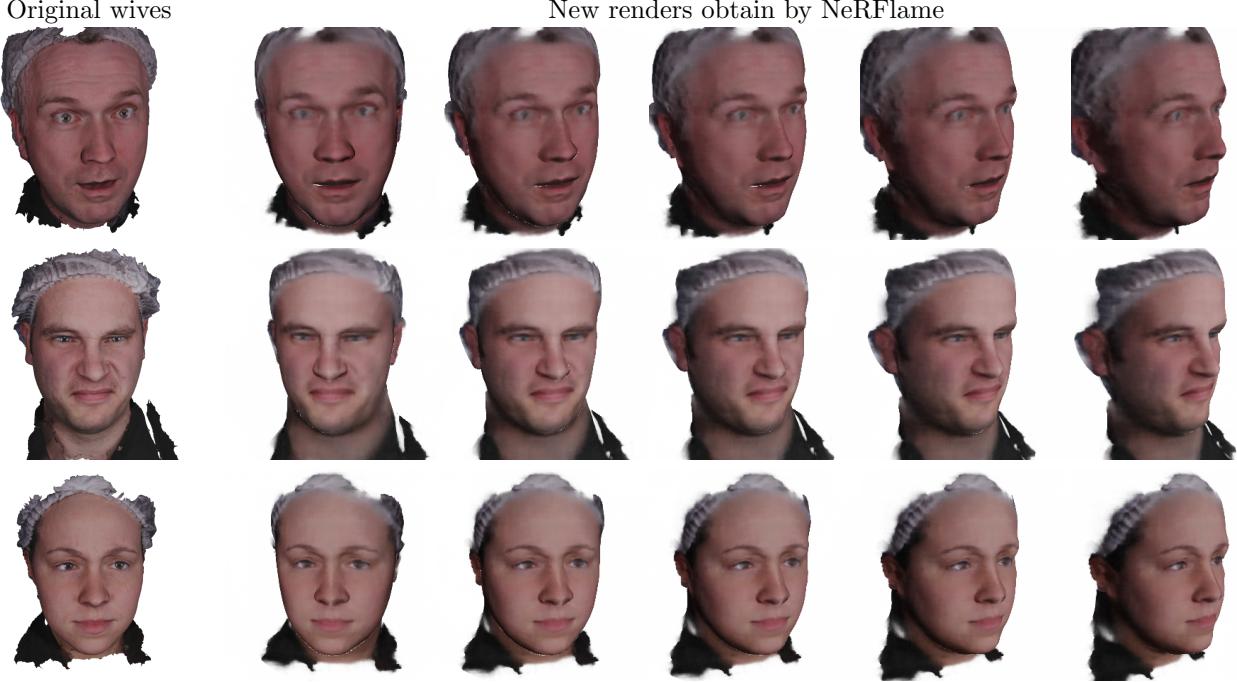
mesh.

**Dynamic Neural Radiance Fields** Current solutions for implicit rephrasing of human face avatars are trained in movies. We assume that we have external tools for segmenting frames in the movie. We often use additional information like each frame’s camera angle or FLAME representation.

In [16], authors implicitly model the facial expressions by conditioning the NeRF with the global expression code obtained from 3DMM tracking [37]. In [43], authors leverage the idea of dynamic neural radiance fields to improve the mouth region’s rendering, which is not represented by the face model motion prior. The IMAvatar [42] model learns the subject-specific implicit representation of texture together with expression. In [17] authors use neural graphics primitives, where for each of the blendshape, a multi-resolution grid is trained. In RigNeRF [5] author propose a model which changes head pose and facial expressions using a deformation field that is guided by a 3D morphable face model (3DMM).

In contrast to the above method, some authors use explicit apprehension. In [2], authors proposed ClipFace, an approach to text-guided editing of textured 3D face models. In [23], we have a one-shot mesh-based model reconstruction. In [44, 12], authors propose a model based on a mixture of 2D and 3D datasets.

Our method is situated between the above approaches. Similarly to the explicit approach, we use a single 3D object instead of movies to train. We also use FLAME mesh to edit the avatar’s shape and expressions. On the other hand, we use an implicit represen-



Rysunek 4. Reconstruction of 3D object obtained by NeRFlame. As we can see, NeRFlame model the detailed appearance of the 3D face.

tation of the colors of objects.

### 3. NeRFlame: FLAME-based conditioning of NeRF for 3D face rendering

In this subsection, we introduce NeRFlame - the novel 3D face representation that combines the benefits of Flame and NeRF models. We first provide the details about the FLAME and NeRF approaches and further describe the concept of NeRFlame and how it can be used for controlled face manipulation.

#### 3.1. FLAME

FLAME (Faces Learned with an Articulated Model and Expressions) [25] is a 3D facial model trained from thousands of accurately aligned 3D scans. The model is factored in that it separates the representation of identity, pose, and facial expression, similar to the human body approach. It is represented by low polygon count, articulation, and blend skinning that is computationally efficient, compatible with existing game and rendering engines, and simple in order to maintain its practicality. The parameters of the model are trained by optimizing the reconstruction loss, assuming detailed temporal registration of our template mesh with three unconnected components, including the base face and two eyeballs.

Formally, the FLAME is a function from human face parametrization  $\mathcal{F}_{flame}(\beta, \psi, \phi)$  where  $\beta, \psi$  and  $\phi$

describe shape, expression, and pose parameters to a mesh with  $n$  vertices:

$$\mathcal{F}_{flame}(\beta, \psi, \phi) : \mathbb{R}^{k_\beta \times k_\psi \times k_\phi} \rightarrow \mathbb{R}^{n \times 3}, \quad (1)$$

where  $k_\beta$ ,  $k_\psi$ , and  $k_\phi$  are the numbers of shape, expression, and pose parameters.

In the classical version, we can fit our model to 3D scans or 2D images by using facial landmarks. Many strategies exist to choose landmarks and parameters for its training [25]. But in high level of generalization for input image 3D scan, (or 2D image)  $I$  and arbitrary chose method for estimation facial landmark points  $\mathcal{LP}$  we minimize L2 distance

$$\min_{(\beta, \psi, \phi)} \|\mathcal{LP}(\mathcal{F}_{flame}(\beta, \psi, \phi)) - \mathcal{LP}(I)\|_2$$

Such an approach is effective, but there are a few limitations. Localizing landmarks and choosing which parameters to optimize first is not trivial. On the other hand, for 2D images, the results are not well-qualified. We can use a pre-trained auto-encoder-based model DECA [14] for face reconstruction from 2D images to solve such problems.

In this paper, we train the FLAME-based model in NeRF training scenario. As input, we take a few 2D images. As an effect, we obtain a correctly fitted FLAME model and NeRF rendering model for now views.

### 3.2. NeRF

NeRF representation of 3D objects NeRFs [29] is the model for representing complex 3D scenes using neural architectures. In order to do that NeRFs take a 5D coordinate as input, which includes the spatial location  $\mathbf{x} = (x, y, z)$  and viewing direction  $\mathbf{d} = (\theta, \psi)$  and returns emitted color  $\mathbf{c} = (r, g, b)$  and volume density  $\sigma$ .

A classical NeRF uses a set of images for training. In such a scenario, we produce many rays traversing through the image and a 3D object represented by a neural network. NeRF approximates this 3D object with an MLP network:

$$\mathcal{F}_{NeRF}(\mathbf{x}, \mathbf{d}; \Theta) = (\mathbf{c}, \sigma).$$

The model is parameterized by  $\Theta$  and trained to map each input 5D coordinate to its corresponding volume density and directional emitted color.

The loss of NeRF is inspired by classical volume rendering [20]. We render the color of all rays passing through the scene. The volume density  $\sigma(\mathbf{x})$  can be interpreted as the differential probability of a ray. The expected color  $C(\mathbf{r})$  of camera ray  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  (where  $\mathbf{o}$  is ray origin and  $\mathbf{d}$  is direction) can be computed with an integral.

In practice, this continuous integral is numerically estimated using a quadrature. We use a stratified sampling approach where we partition our ray  $[t_n, t_f]$  into  $N$  evenly-spaced bins and then draw one sample  $t_i$  uniformly at random from within each bin. We use these samples to estimate  $C(\mathbf{r})$  with the quadrature rule [28]:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i,$$

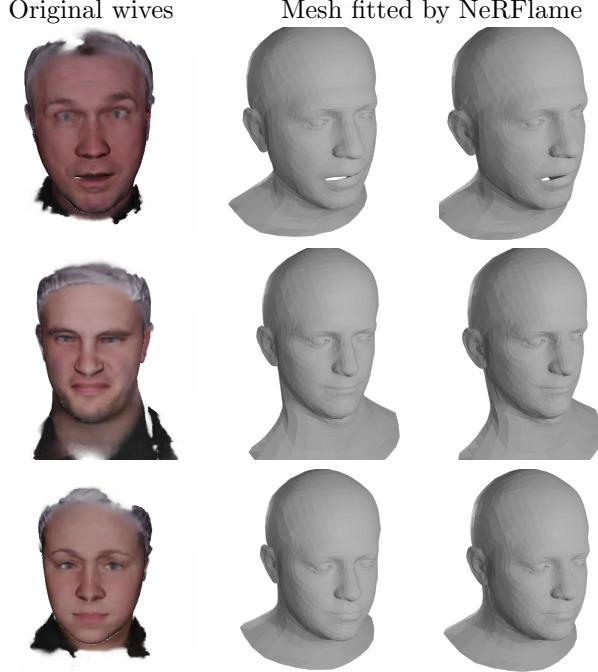
$$\text{where } T_i = \exp \left( - \sum_{j=1}^{i-1} \sigma_j \delta_j \right),$$

where  $\delta_i = t_{i+1} - t_i$  is the distance between adjacent samples. This function for calculating  $\hat{C}(\mathbf{r})$  from the set of  $(\mathbf{c}_i, \sigma_i)$  values is trivially differentiable.

We then use the volume rendering procedure to render the color of each ray from both sets of samples. Contrary to the baseline NeRF [29], where two "coarse" and "fine" models were simultaneously trained, we use only the "coarse" architecture. Our loss is simply the total squared error between the rendered and true pixel colors

$$= \sum_{\mathbf{r} \in R} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_2^2 \quad (2)$$

where  $R$  is the set of rays in each batch, and  $C(\mathbf{r})$ ,  $\hat{C}(\mathbf{r})$  are the ground truth and predicted RGB colors for ray  $\mathbf{r}$  respectively.



Rysunek 5. During the training of NeRFlame, we simultaneously model FLAME mesh and NeRF dedicated to colors. In the above figure, we present the meshes fitted by NeRFlame.

### 3.3. NeRFlame

We introduce NeRFlame the 3D face model that combines the benefits of mesh representations from FLAME and NeRF implicit rephrasing of 3D objects. Thanks to the application of NeRFs, we are able to estimate the parameters of FLAME directly from 2D images without using landmarks points. On the other hand, we obtain NeRF model, which can be edited similarly to FLAME. In order to achieve that, we introduce the function that approximates the volume density using FLAME model.

Consider the distance function  $d(\mathbf{x}, \mathcal{M})$  between point  $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$  and the mesh  $\mathcal{M} := \mathcal{M}_{\beta, \psi, \phi}$  created by  $F_{flame}(\beta, \psi, \phi)$ , with parameters  $\beta, \psi, \phi$ . Note, that edges between vertices in FLAME model can be directly taken from the template mesh (see [4] for details). We define the volume density function as:

$$\sigma(\mathbf{x}, \mathcal{M}) = \begin{cases} 0, & \text{if } d(\mathbf{x}, \mathcal{M}) > \varepsilon \\ 1 - \frac{1}{\varepsilon} d(\mathbf{x}, \mathcal{M}), & \text{otherwise,} \end{cases} \quad (3)$$

where  $\varepsilon$  is a hyperparameter that defines the neighborhood of the mesh surface. In practice, the values of the density volume function are non-zero only in the close neighborhood of the mesh.

The NeRFlame can be represented by the function:

$$\mathcal{F}_{NeRFlame}(\mathbf{x}; \beta, \psi, \phi, \Theta) = (\mathcal{F}_\Theta(\mathbf{x}), \sigma(\mathbf{x}, \mathcal{M})), \quad (4)$$

where  $\mathcal{F}_\Theta(\mathbf{x})$  is the MLP that predicts the color, similar to the NeRF model.

The model is trained in an end-to-end manner directly, optimizing the criterion (2) with respect to the MLP parameters  $\Theta$ , and FLAME parameters  $\beta, \psi, \phi$ , which describe shape, expression, and pose. In NeRFlame, we utilize the original loss function used to train NeRF models. Therefore, the structure of colors on rays must be consistent. During training, the model modified the mesh structure to be consistent with the 3D structure of the target face. Simultaneously neural network  $\mathcal{F}$  produces colors for the rendering procedure.

### 3.4. Controlling NeRF models to obtain face manipulation

The classical NeRF method is known to generate highly detailed and realistic images. However, it can be challenging to manipulate NeRF models to achieve precise facial modifications. Several techniques, such as generative models, dynamic scene encoding, and conditioning mechanisms, have been proposed to address this challenge. Nonetheless, controlling NeRF models to the same extent as mesh representations remains elusive. In contrast, the FLAME is a straightforward model with three parameters, namely  $\beta, \psi$ , and  $\phi$ , representing shape, expression, and pose, respectively. By performing simple linear operations on these parameters, it is possible to rotate the avatar, change facial expressions, and adjust facial features to a certain degree.

Our NeRFlame is built on the FLAME model so that we can manipulate our FLAME model to control density prediction  $\sigma$ . But the problem is with predicting RGB colors after transformation. To solve such a problem, we use a simple technique. For the prediction of color after transformation, we will return to the starting position where the color is known, see Fig. 6.

Let us consider NeRFlame model, which is already trained. We have parameters  $\beta_1, \psi_1, \phi_1, \Theta, \varepsilon$ , function

$$\mathcal{F}_{NeRFlame}(\mathbf{x}; \beta_1, \psi_1, \phi_1, \Theta)$$

and fitted mesh  $\mathcal{M}_1$ , created by the FLAME from parameters  $\beta_1, \psi_1, \phi_1$ . Let us assume that we apply some modification of FLAME parameters, which means that we obtain new  $\beta_2, \psi_2, \phi_2$ . Using these parameters, we can create the modified mesh  $\mathcal{M}_2$ , simply using the FLAME model. Instead of retraining the NeRF model for new parameters requiring the 2D images for a new pose, we propose applying a simple transformation between the modified and original space.

We consider some point  $\mathbf{x}_2$  located on mesh  $\mathcal{M}_2$  representing the new pose. We postulate using the affine transformation  $T(\cdot)$ , which transforms the point on the mesh to the original pose, in which the model was trained:

$$T(\mathbf{x}_2) = \mathbf{x}_1, \quad (5)$$

where  $\mathbf{x}_1 \in \mathcal{M}_1$  is the corresponding point on original pose. We can identify an element in the original pose for each element on the mesh after the transformation. In practice, finding the transformation  $T$  is not trivial since it depends on the local transformation of the mesh.

However, for a given point  $\mathbf{x}_1 \in \mathcal{M}_1$ , we can find a face triangle described by vertices  $(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3) \in V_1$  that contains  $\mathbf{x}_1$ , where  $V_1$  is the set of vertices for mesh  $\mathcal{M}_1$ . Because FLAME model is shifting the vertices of the model keeping the connections unchanged, we can locate the corresponding triangle  $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3) \in V_2$  in mesh  $\mathcal{M}_2$ . For each of the triangles that create the mesh, we define affine transformation:

$$T_m(\mathbf{p}_i) = \mathbf{q}_i, \text{ for } i = 1, 2, 3. \quad (6)$$

In such a situation, we assume that such transformations  $T_m(\cdot)$  are affine, and we can use Least-Squares Conformal Multilinear Regression [33] to estimate the parameters. Practically, finding the transformation for each of the triangles is extremely fast, requires inverting a fourth-dimensional matrix, and can be parallelized. Having the parameters of transformations estimated, we can apply them directly to the  $\mathbf{x}$  in NeRFlame model given by (4) and calculate the colors as in an unshifted pose.

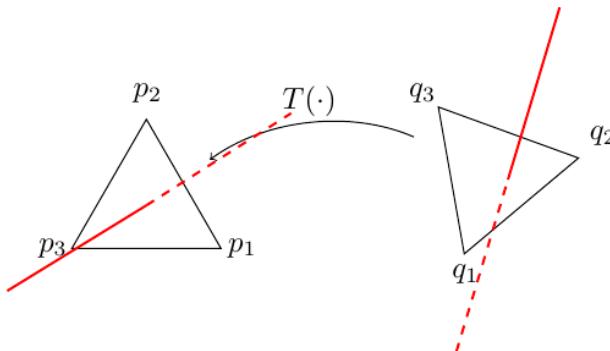
## 4. Experiments

In this section, we describe the experimental results of the proposed model. To our knowledge, it is the first model that obtains editable NeRF trains on a single object in one position. Most of the methods use movies to encode many different positions of the face. Therefore, it is hard to compare our results to other algorithms. In the first subsection, we show that our model produces high-quality NeRF representations of the objects by comparing our model with our baseline classical NeRF and classical textured FLAME. In the second subsection, we present meshes obtained by our model. Finally, we show that our model allows facing manipulations.

Since the current literature does not provide suitable data sets for evaluating the NeRF-based model for modeling 3D face avatars, we created a data set using 3D scenes. We create a classical NeRF training data

	PSNR $\uparrow$			SSIM $\uparrow$			LPIPS $\downarrow$		
	NeRF	NeRFlame	FLAME	NeRF	NeRFlame	FLAME	NeRF	NeRFlame	FLAME
Face 1	33.37	20.99	9.67	0.96	0.88	0.76	0.05	0.14	0.26
Face 2	33.08	22.89	12.97	0.97	0.91	0.82	0.04	0.10	0.20
Face 3	31.96	21.35	12.51	0.96	0.88	0.79	0.04	0.13	0.23
Face 4	33.39	22.14	12.44	0.96	0.89	0.80	0.05	0.14	0.24
Face 5	33.15	23.04	11.30	0.96	0.89	0.77	0.05	0.11	0.26
Face 6	32.40	21.47	11.22	0.95	0.86	0.76	0.06	0.15	0.26
Face 7	32.42	22.74	11.45	0.96	0.89	0.76	0.06	0.13	0.26

Tabela 1. Comparison of PSNR, SSIM, and LPIPS matrices between our model and NeRF and FLAME baselines. As we can see NeRF gives essentially better since do not allow manipulation. On the other hand, FLAME model gives inferior results since we train mesh and texture only on landmark points.



Rysunek 6. Visualization of transformation in NeRFlame. We aim to aggregate colors along the ray during rendering in the new position (see the red line in the right image). NeRFlame uses FLAME mesh, therefore we can localize the face's vertex, which is crossed with the ray  $p_1, p_2, p_3$  and the corresponding triangle in the initial position mesh  $q_1, q_2, q_3$ . Thanks to such pairs of points, we estimate affine transformation  $T$ , which is used to find the ray in the initial position (see the red line in the left image).

set. We construct  $200 \times 200$  transparent background images from random camera positions.

#### 4.1. Reconstruction Quality

In this subsection, we show that NeRFlame can reconstruct a 3D human face with similar quality as classical NeRF. Since we train our model on a single position, it is difficult to compare our model to dynamic neural radiance fields. Therefore we show that our model has a slightly lower quality than classical NeRF but allows dynamic modification. On the other hand, we show that we obtain better results than textured FLAME, which is not able to capture the geometry and appearance details of the human face.

In Fig 7, we compare NeRFlame and textured FLA-

ME. As we can see, NeRFlame can reproduce facial features and geometry. On the other hand, FLAME produces well-suited textures, but the mesh is not well-suited. In Tab. 1, we present numerical cooperation. We compare the metric reported by NeRF called PSNR (peak signal-to-noise ratio), SSIM (structural similarity index measure), LPIPS (learned perceptual image patch similarity) used to measure image reconstruction effectiveness. As we can see NeRF gives essentially better since do not allow manipulation. On the other hand, FLAME model gives inferior results since we train mesh and texture only on landmark points. In Fig. 4, we present new renders of the model obtained by NeRFlame. As we can see, NeRFlame model the detailed appearance of the 3D face.

#### 4.2. Mesh fitting

The RGB colors generated by NeRF are present only in the  $\varepsilon$  vicinity of the mesh. This approach allows for a precise fitting of the mesh to the human face, which is critical for generating animated models. In Figure 5, we present the rendered faces and corresponding meshes produced by our NeRFlame approach. The results demonstrate that our method can accurately capture the underlying mesh structure.

#### 4.3. Face manipulation

Our approach enables the manipulation of human facial features. Leveraging FLAME as a backbone, NeRFlame offers the ability to manipulate FLAME features and modify NeRF representations. In Figure 7, we showcase three facial expressions, including open mouths and head rotations, that can be manipulated using our model in a manner similar to the classical FLAME model.

NeRFlame simultaneously train mesh and NeRF components for color. After training, we can exchange

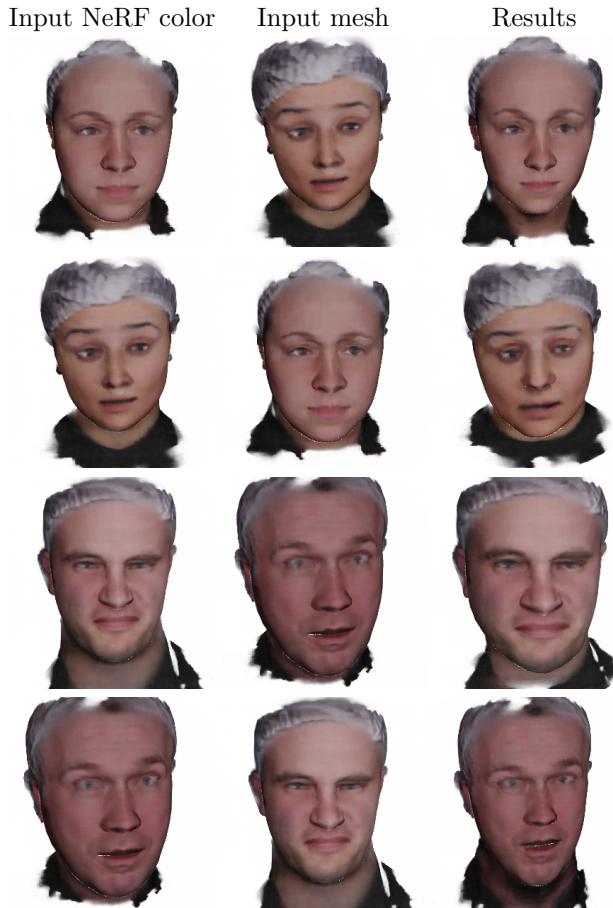


Rysunek 7. Our model allows producing manipulation of the human face. In NeRFFlame, we use FLAME for volume density rendering. Therefore, we can manipulate FLAME features and modify NeRF representation. In the figure, we show three faces and its version with open mouths and head rotations.

the produced mesh in our model to obtain a modification of the final look of the avatar, see Fig. 8.

## 5. Conclusions

In this work, we introduce a novel approach called NeRFlame, which combines NeRF and FLAME to achieve high-quality rendering and precise pose control. While NeRF-based models use neural networks to model RGB colors and volume density, our method utilizes an explicit density volume represented by the FLAME mesh. This allows us to model the quality of NeRF rendering and accurately control the appearance of the final output. As a result of offering complete control over the model, the quantitative performance of our approach is marginally inferior to that of a static NeRF model.



Rysunek 8. NeRFlame simultaneously train mesh and NeRF components for color. After training, we have two components: FLAME-base mesh and NeRF-based colors. We can produce mixed models by taking FLAME from one model and NeRF from a different one. In consequence, we obtain an avatar that takes face geometry from the first model and colors from the second one.

## Literatura

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In International conference on machine learning, pages 40–49. PMLR, 2018.
- [2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. arXiv preprint arXiv:2212.01406, 2022.
- [3] Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D Kulkarni, and Joshua B Tenenbaum. Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1511–1519, 2017.
- [4] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. arXiv preprint arXiv:2108.04913, 2021.
- [5] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20364–20373, 2022.
- [6] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgbd surface reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6290–6301, 2022.
- [7] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5855–5864, 2021.
- [8] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5470–5479, 2022.
- [9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khassis, et al. Efficient geometry-aware 3d generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16123–16133, 2022.
- [10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII, pages 333–350. Springer, 2022.
- [11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruc-

- tion. In European conference on computer vision, pages 628–644. Springer, 2016.
- [12] Radek Daněček, Michael Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20279–20290. IEEE, 2022.
- [13] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12882–12891, 2022.
- [14] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG), 40(4):1–13, 2021.
- [15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qin-hong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5501–5510, 2022.
- [16] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8649–8658, 2021.
- [17] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. ACM Transactions on Graphics (TOG), 41(6):1–12, 2022.
- [18] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In European Conference on Computer Vision, pages 484–499. Springer, 2016.
- [19] Christian Häne, Shubham Tulsiani, and Jitendra Malik. Hierarchical surface prediction for 3d object reconstruction. In 2017 International Conference on 3D Vision (3DV), pages 412–420. IEEE, 2017.
- [20] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. ACM SIGGRAPH computer graphics, 18(3):165–174, 1984.
- [21] Adam Kania, Artur Kasymov, Maciej Zieba, and Przemysław Spurek. Hypernerfgan: Hypernetwork approach to 3d nerf gan. arXiv preprint arXiv:2301.11631, 2023.
- [22] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzcinski, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18602–18611. IEEE, 2022.
- [23] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II, pages 345–362. Springer, 2022.
- [24] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. ACM Transactions on Graphics (TOG), 36(4):1–14, 2017.
- [25] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph., 36(6):194–1, 2017.
- [26] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. Advances in Neural Information Processing Systems, 33:15651–15663, 2020.
- [27] Zehua Liu, Yuhe Zhang, Jian Gao, and Shurui Wang. Vfmvac: View-filtering-based multi-view aggregating convolution for 3d shape recognition and retrieval. Pattern Recognition, 129:108774, 2022.
- [28] Nelson Max. Optical models for direct volume rendering. IEEE Transactions on Visualization and Computer Graphics, 1(2):99–108, 1995.
- [29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.
- [30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG), 41(4):1–15, 2022.
- [31] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5480–5490, 2022.
- [32] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12892–12901, 2022.
- [33] Kendra K Schmid, David B Marx, and Ashok Samal. Tridimensional regression for comparing and mapping 3d anatomical structures. Anatomy Research International, 2012, 2012.
- [34] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. Wasserstein distributional harvesting for highly dense 3d point clouds. Pattern Recognition, 132:108978, 2022.
- [35] Ayan Sinha, Jing Bai, and Karthik Ramani. Deep learning 3d shape surfaces using geometry images. In European Conference on Computer Vision, pages 223–240. Springer, 2016.
- [36] Matthew Tancik, Vincent Casser, Xinchen Yan, Sa-beek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8248–8258, 2022.
- [37] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2387–2395. IEEE Computer Society, 2016.
  - [38] T Trzciński. Points2nerf: Generating neural radiance fields from 3d point cloud. arXiv preprint arXiv:2206.01290, 2022.
  - [39] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Refnerf: Structured view-dependent appearance for neural radiance fields. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5481–5490. IEEE, 2022.
  - [40] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5610–5619, 2021.
  - [41] Fei Yang, Franck Davoine, Huan Wang, and Zhong Jin. Continuous conditional random field convolution for point cloud segmentation. Pattern Recognition, 122:108357, 2022.
  - [42] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13535–13545. IEEE, 2022.
  - [43] Wojciech Zienonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. arXiv preprint arXiv:2211.12499, 2022.
  - [44] Wojciech Zienonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII, pages 250–269. Springer, 2022.