FLNeRF: 3D Facial Landmarks Estimation in Neural Radiance Fields Supplementary Material

Anonymous authors
Paper under double-blind review

Please watch the supplementary video for dynamic 3D face visualization.

1 Training Details

We train the coarse and fine models one after the other. First, the coarse model is trained on the augmented data set as described in Section 4.1. Then, the well-trained coarse model predicts the transform matrix and coarse landmarks which locate the 4 sampling positions for the fine model. Together with the expression augmentation Section 4.2, the training data set for fine model is generated.

To balance training speed and sampling fidelity, the resolution of NeRF sampling box in the coarse and fine models are respectively $64 \times 64 \times 64$. We select 100 identities from FaceScape dataset to perform data augmentation for coarse and fine models. Furthermore, 5 extra identities are randomly chosen for testing with data augmentation.

When training the coarse and fine model, we set learning rate to 0.001 and batch size 32. The learning rate will finally decay to 8e-6. We train our coarse model for 100 epochs and fine model for 50 epochs. We train our FLNeRF on 4x GTX 1080 GPUs. Training coarse model takes around 8 hours and fine model takes around 12 hours.

2 More about Comparisons

2.1 More metrics

Table 1 shows quantitative comparison of our FLNeRF with state-of-the-art methods in terms of average adaptive Wing loss (Wang et al., 2020). For hyperparameters of adaptive Wing loss, we adopt default values given in the official implementation code: $\omega = 14$, $\theta = 0.5$, $\epsilon = 1$, and $\alpha = 2.1$. Table 2 shows quantitative comparison of our FLNeRF with state-of-the-art methods in terms of average MSE. These two tables and Table 1 in our main paper show that our FLNeRF perform significantly better than all representative works.

Table 1: Quantitative comparison of FLNeRF and representative methods in average adaptive Wing loss. All values are multiplied by 10^3. Empty entries mean different landmarks definition on the corresponding regions.

Method	Predictor	Average Adaptive Wing Loss of All Expressions			Average A.W.L. of Exaggerated Expression			Avg. #Fail
		Mouth	Eyes	Nose	Mouth	Eyes	Nose	
2D Estimation + Triangulation	RSN	3.39±0.86	-	-	2.45±0.45	-	-	0.00
	RTMDet	3.39±0.79	-	3.78±0.45	2.46±0.42	-	3.86±0.41	0.00
	DarkPose	3.46±0.84	-	4.09±0.53	2.51±0.52	-	4.32±0.61	0.00
	DeepPose-SW	3.53±0.95	-	4.04±0.54	2.58±0.49	-	4.44±0.57	0.00
	STAR	32.12±104.2	3.82±6.08	39.35±118.46	164.24±189.48	18.80±22.38	331.49±368.18	56.83
	2D FAN	3.29±0.74	2.45±0.18	3.79±0.28	2.32±0.30	2.52±0.09	3.90±0.30	5.02
	SPIGA	22.66±82.58	2.88±0.43	22.94±63.72	84.34±135.95	3.11±0.45	124.99±181.49	57.99
	PIPNet	4.11±0.99	2.84±0.42	4.48±0.55	3.00±0.58	3.11±0.50	4.71±0.54	2.11
Averaged 3D Estimation on Single Images	3DDFA	1.10±0.82	0.82±0.25	1.08±0.43	3.13±0.74	1.11±0.27	1.27±0.51	48.29
	SynergyNet	1.27±0.87	2.39±0.47	1.13±0.21	0.87±0.33	2.73±0.47	1.12±0.23	1.88
	3D FAN	1.13±0.82	2.11±0.51	1.10±0.27	0.72±0.27	2.88±0.69	1.01±0.32	2.00
	DECA	1.10±0.81	1.71±0.32	0.85±0.19	0.80±0.23	2.25±0.33	0.88±0.12	0.00
Estimation on NeRF	**FLNeRF**	**0.54±0.86**	**0.17±0.08**	**0.22±0.22**	**0.18±0.12**	**0.21±0.04**	**0.14±0.06**	-

Table 2: Quantitative comparison of FLNeRF and representative methods in average MSE. All values are multiplied by 10^2. Empty entries mean different landmarks definition on the corresponding regions.

Method	Predictor	Average MSE of All Expressions			Average MSE of Exaggerated Expression			Avg. #Fail
		Mouth	Eyes	Nose	Mouth	Eyes	Nose	
2D Estimation + Triangulation	RSN	1.83±0.49	-	-	1.63±0.37	-	-	0.00
	RTMDet	1.75±0.39	-	1.65±0.33	1.57±0.26	-	1.71±0.21	0.00
	DarkPose	1.81±0.42	-	1.95±0.43	1.61±0.48	-	2.21±0.50	0.00
	DeepPose-SW	1.87±0.56	-	1.88±0.43	1.64±0.38	-	2.32±0.51	0.00
	STAR	3.85±1.44	2.86±0.83	3.54±1.43	4.51±1.76	4.05±1.51	6.01±2.82	56.83
	2D FAN	1.63±0.30	1.58±0.10	1.70±0.10	1.33±0.08	1.59±0.07	1.78±0.12	5.02
	SPIGA	3.46±1.35	2.51±0.85	1.24±3.18	3.68±1.29	3.71±1.37	5.08±2.24	57.99
	PIPNet	2.85±1.01	2.15±0.58	2.44±0.65	2.35±0.72	2.34±0.61	2.83±0.67	2.11
Averaged 3D Estimation on Single Images	3DDFA	0.21±0.21	0.28±0.08	0.13±0.03	1.05±0.13	0.38±0.10	0.15±0.03	48.29
	SynergyNet	0.43±0.25	0.84±0.10	0.23±0.04	1.08±0.28	1.03±0.10	0.24±0.04	1.88
	3D FAN	0.38±0.21	0.86±0.13	0.25±0.05	0.82±0.21	1.16±0.18	0.22±0.04	2.00
	DECA	0.33±0.21	0.63±0.10	0.13±0.03	0.83±0.19	0.88±0.12	0.14±0.02	0.00
Estimation on NeRF	**FLNeRF**	**0.05±0.08**	**0.02±0.01**	**0.02±0.02**	**0.02±0.01**	**0.02±0.01**	**0.01±0.01**	-

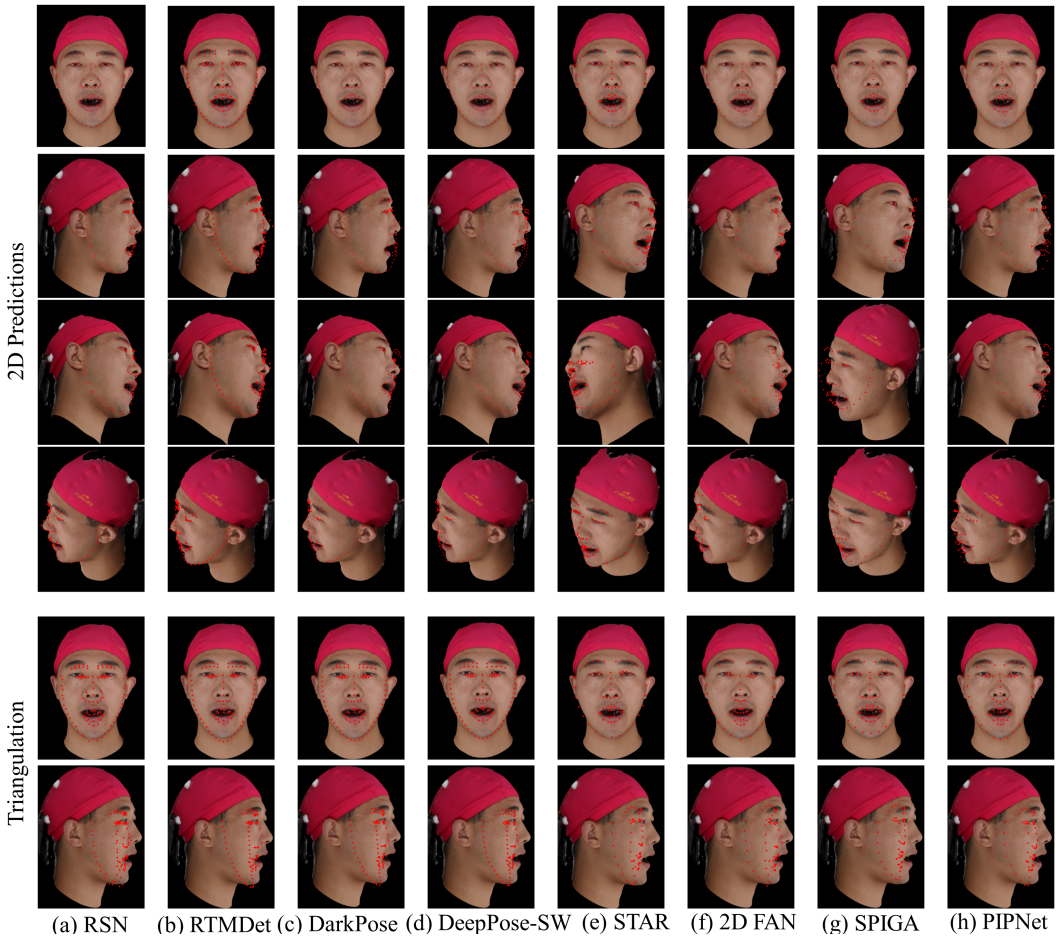

Figure 1: Visualizations of **2D landmarks estimation** by 2D landmarks predictors, and **triangulated 3D landmarks** overlayed on a frontal-view and a lateral-view image. The upper four rows are 2D landmarks prediction results. Note the 2D landmarks do not have depth and the occluded landmarks are highly inaccurate. The bottom two rows show triangulation results overlayed on a frontal view and a side view image. All images on the same row are of the same view directions, except the 2nd to 4th rows of column (e)(g), where SPIGA and STAR malfunction.

2.2 WHY 2D ESTIMATION FOLLOWED BY TRIANGULATION PRODUCE WORSE RESULTS?

Fig. 1 shows visualizations of 2D landmarks estimation by 2D landmarks predictors, and triangulated 3D landmarks overlayed on a frontal view and a lateral view image. All 2D predictors give good approximations on the frontal view image, but their performance varies under different camera poses. In general, images taken near the frontal view are easier for 2D predictors to give more reasonable estimation of 2D landmarks. When the image is taken from lateral views, like the 2nd

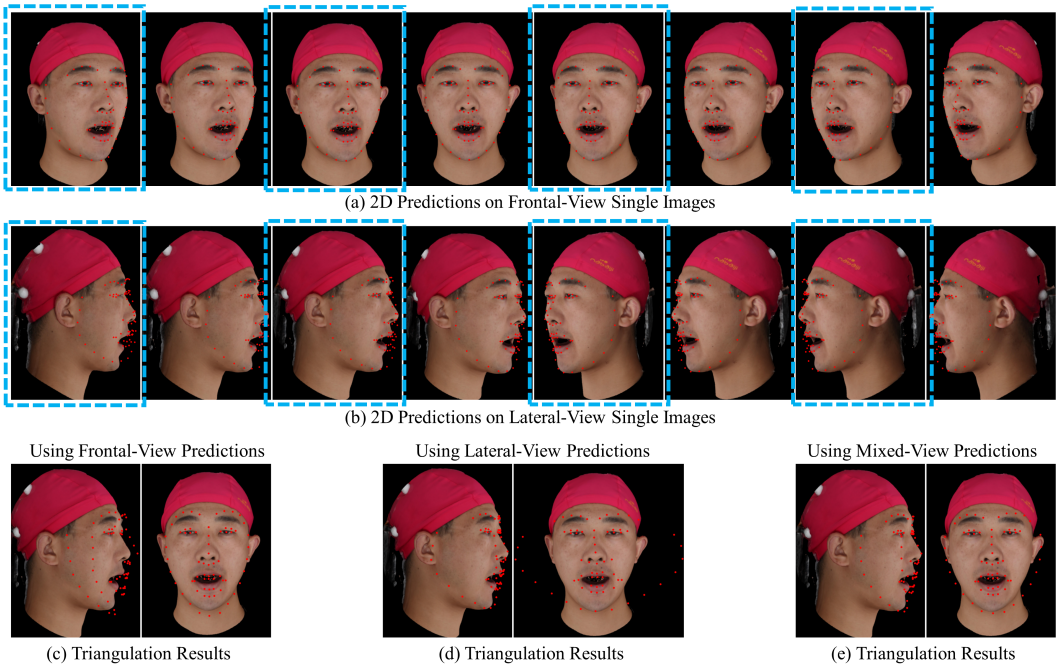

Figure 2: **Triangulation results using estimations on different combinations of 8 single images**. (a) and (b) shows 2D predictions by PIPNet (Jin et al., 2021) on frontal-view and lateral-view images, respectively. (c) demonstrates triangulation results using estimations on frontal-view images shown in (a). (d) demonstrates triangulation results using estimations on lateral-view images shown in (b). (e) demonstrates triangulation results using estimations on mixed-view images, which are framed by blue dashed square boxes in (a)(b).

to 4th rows of Fig. 1, 2D landmark predictions often yield highly inaccurate results for occluded landmarks, as well as landmarks that are located close to occluding boundaries. Triangulation accumulates all these errors thus performing bad.

To further validate our conclusion, we utilize PIPNet (Jin et al., 2021) as the 2D landmarks predictor to assess the impact of selected images on 2D landmark estimation followed by triangulation. Fig. 2 illustrates the 2D landmarks predictions by PIPNet for both frontal-view and lateral-view images, as well as the visualized results of triangulation performed on estimations from frontal-view, lateral-view, and mixed-view images. As depicted in Fig. 2, the 2D estimations on frontal-view images demonstrate superior accuracy in approximating the locations and shapes of the mouth, eyes, nose, and cheeks, but provide limited depth information about the faces. Consequently, the triangulated 3D landmarks derived from frontal-view estimations offer more precise predictions regarding the locations and shapes of the mouth, eyes, nose, and cheeks, albeit resulting in significant inaccuracies in depth estimation. Conversely, the 2D estimations on lateral-view images exhibit better approximation of depth information but lack knowledge about the locations and shapes of the mouth, eyes, nose, and cheeks. As a result, the triangulated 3D landmarks based on lateral-view estimations yield more accurate depth estimation while introducing considerable inaccuracies in the predictions of the locations and shapes of the mouth, eyes, nose, and cheeks. As depicted in Fig. 2 (e), when performing triangulation using a combination of half of the frontal-view estimations and half of the lateral-view images, a more balanced trade-off is achieved between depth estimation and estimations of locations and shapes of the mouth, eyes, nose, and cheeks.

We then investigate into the impact of the number of images used for 2D face landmark estimation on triangulation performance. We perform separate 2D estimations followed by triangulation using 8, 24, and 40 images sampled uniformly from various view directions. To maintain consistent terminology, we refer to these selected images as "mixed-view images". Qualitative results of triangulation on 8, 24, and 40 mixed-view estimations are depicted in Fig. 3. Due to malfunctioning on a significant number of lateral-view images ($\#Fail > 55$), we did not test STAR and SPIGA. It can be observed that the triangulated 3D landmarks using only 8 2D estimations yield sub-optimal approximations, particularly for the landmarks on the cheeks. As the number of input images increases, the landmarks on the cheeks become more closely aligned with the faces; however, the accuracy of depth estimation progressively diminishes. Quantitative results are presented in Table 3, Table 4, and Table 5. From these tables, we can observe a significant deterioration in performance

Table 3: Average Wing loss of triangulation results using 2D estimations on different number of mixed-view images. All values are multiplied by 10. Empty entries mean different landmarks definition on the corresponding regions.

#Images	Predictor	Average Wing Loss of All Expressions			Average Wing Loss of Exaggerated Expression		
		Mouth	Eyes	Nose	Mouth	Eyes	Nose
8	RSN	1.64±0.32	-	-	1.80±0.21	-	-
	RTMDet	1.75±0.29	-	1.64±0.14	2.00±0.31	-	1.60±0.05
	DarkPose	2.03±0.88	-	1.84±0.47	3.52±1.43	-	2.49±0.84
	DeepPose-SW	1.65±0.41	-	1.48±0.38	2.02±0.24	-	1.18±0.21
	2D FAN	1.46±0.62	1.41±0.42	1.76±0.35	1.47±0.32	1.26±0.13	1.81±0.05
	PIPNet	1.48±0.37	1.64±0.21	1.26±0.15	1.46±0.22	1.52±0.21	1.39±0.12
24	RSN	1.65±0.36	-	-	1.88±0.24	-	-
	RTMDet	1.62±0.38	-	0.93±0.12	1.87±0.26	-	0.83±0.07
	DarkPose	1.80±0.65	-	1.22±0.43	2.33±0.64	-	1.98±0.84
	DeepPose-SW	1.84±0.49	-	1.43±0.27	2.33±0.19	-	2.06±0.28
	2D FAN	1.75±0.42	2.06±0.24	1.70±0.23	1.72±0.21	2.00±0.12	1.78±0.16
	PIPNet	2.08±0.72	1.86±0.27	1.38±0.40	1.97±0.55	1.87±0.25	1.70±0.35
40	RSN	3.27±0.38	-	-	3.38±0.43	-	-
	RTMDet	3.21±0.38	-	2.86±0.13	3.32±0.27	-	2.94±0.22
	DarkPose	3.34±0.58	-	3.22±0.35	3.44±0.90	-	3.71±0.73
	DeepPose-SW	3.62±0.54	-	3.38±0.29	3.81±0.32	-	4.03±0.39
	2D FAN	3.61±0.42	4.00±0.19	3.69±0.18	3.38±0.29	3.99±0.11	3.82±0.22
	PIPNet	4.26±0.84	3.76±0.44	3.69±0.57	3.96±0.67	3.87±0.48	4.05±0.55

Table 4: Average adaptive Wing loss of triangulation results using 2D estimations on different number of mixed-view images. All values are multiplied by 10^3. Empty entries mean different landmarks definition on the corresponding regions.

#Images	Predictor	Average Adaptive Wing Loss of All Expressions			Average A.W.L. of Exaggerated Expression		
		Mouth	Eyes	Nose	Mouth	Eyes	Nose
8	RSN	0.78±0.40	-	-	0.51±0.10	-	-
	RTMDet	0.95±0.40	-	2.32±0.46	0.82±0.52	-	2.70±0.36
	DarkPose	1.20±0.94	-	2.20±0.52	2.60±1.93	-	2.50±0.49
	DeepPose-SW	0.79±0.51	-	1.23±0.90	0.70±0.35	-	0.53±0.20
	2D FAN	0.62±0.60	0.74±0.34	1.97±0.45	0.42±0.31	0.63±0.16	1.99±0.53
	PIPNet	0.71±0.41	0.84±0.24	1.26±0.33	0.53±0.31	0.79±0.20	1.25±0.26
24	RSN	0.64±0.25	-	-	0.61±0.14	-	-
	RTMDet	0.59±0.27	-	0.18±0.09	0.57±0.15	-	0.14±0.02
	DarkPose	0.79±0.68	-	0.45±0.53	1.05±0.47	-	1.36±1.13
	DeepPose-SW	0.86±0.55	-	0.68±0.39	0.99±0.24	-	1.51±0.30
	2D FAN	0.68±0.39	1.07±0.28	0.62±0.31	0.47±0.11	1.02±0.13	0.63±0.09
	PIPNet	1.27±0.91	1.89±0.25	1.09±0.48	0.90±0.45	0.91±0.19	1.30±0.38
40	RSN	2.36±0.58	-	-	1.81±0.36	-	-
	RTMDet	2.31±0.60	-	2.33±0.24	1.73±0.34	-	2.45±0.41
	DarkPose	2.41±0.77	-	2.69±0.45	1.85±0.71	-	3.18±0.92
	DeepPose-SW	2.65±0.86	-	2.84±0.44	2.08±0.38	-	3.54±0.51
	2D FAN	2.55±0.66	2.16±0.26	2.80±0.29	1.78±0.33	2.16±0.15	2.90±0.39
	PIPNet	3.37±1.07	1.98±0.36	3.54±0.61	2.40±0.63	2.05±0.38	3.69±0.55

Table 5: Average MSE of triangulation results using 2D estimations on different number of mixed-view images. All values are multiplied by 10^2. Empty entries mean different landmarks definition on the corresponding regions.

#Images	Predictor	Average MSE of All Expressions			Average MSE of Exaggerated Expression		
		Mouth	Eyes	Nose	Mouth	Eyes	Nose
8	RSN	0.17±0.10	-	-	0.20±0.05	-	-
	RTMDet	0.19±0.09	-	0.17±0.03	0.25±0.08	-	0.18±0.01
	DarkPose	0.36±0.57	-	0.26±0.30	1.14±0.96	-	0.66±0.58
	DeepPose-SW	0.19±0.09	-	0.13±0.07	0.25±0.05	-	0.09±0.03
	2D FAN	0.16±0.34	0.15±0.25	0.21±0.22	0.13±0.05	0.11±0.02	0.20±0.02
	PIPNet	0.15±0.10	0.16±0.04	0.12±0.03	0.15±0.05	0.15±0.03	0.14±0.02
24	RSN	0.17±0.10	-	-	0.23±0.05	-	-
	RTMDet	0.17±0.10	-	0.05±0.01	0.23±0.07	-	0.04±0.01
	DarkPose	0.25±0.30	-	0.11±0.16	0.40±0.28	-	0.35±0.31
	DeepPose-SW	0.23±0.13	-	0.12±0.06	0.36±0.07	-	0.27±0.07
	2D FAN	0.18±0.14	0.25±0.09	0.16±0.07	0.17±0.04	0.24±0.03	0.17±0.03
	PIPNet	0.35±0.28	0.26±0.11	0.17±0.11	0.27±0.13	0.28±0.10	0.23±0.10
40	RSN	0.77±0.17	-	-	0.81±0.24	-	-
	RTMDet	0.75±0.17	-	0.54±0.06	0.75±0.12	-	0.60±0.12
	DarkPose	0.84±0.38	-	0.73±0.25	0.94±0.55	-	1.09±0.54
	DeepPose-SW	1.00±0.36	-	0.81±0.21	1.07±0.20	-	1.28±0.28
	2D FAN	0.91±0.23	1.02±0.10	0.84±0.12	0.75±0.12	1.01±0.05	0.89±0.12
	PIPNet	1.66±0.82	1.15±0.42	1.31±0.49	1.31±0.51	1.18±0.40	1.44±0.40

for all methods when the number of input images reaches 40. Therefore, having a larger number of available multi-view images does not necessarily imply improved triangulation performance. Even with a smaller number of input images, the performance of the triangulation method is still significantly worse compared to our FLNeRF method. The reason behind this disparity lies in the inability of 2D estimation to capture 3D information, leading to minor inaccuracies that accumulate and re-

Figure 3: **Triangulation results using estimations on different number of mixed-view images**.

sult in significant errors. This highlights the superiority of our method, as direct estimation on NeRF allows for better capture of multi-view 3D information.

2.3 WHY AVERAGED 3D ESTIMATION ON SINGLE IMAGES PRODUCE WORSE RESULTS?

Similar to the reason why the 2D estimation followed by triangulation is not accurate, the 3D landmarks obtained from frontal views achieve precise shape prediction for the mouth, eyes, nose, and cheeks. However, they exhibit inaccurate depth estimation for facial profiles. On the other hand, results derived from lateral views provide accurate depth estimation for the facial profile but yield poor approximation of the locations and shapes of the mouth, eyes, nose, and cheeks. This issue is illustrated in Fig. 4. By averaging the individually calculated 3D estimation results from these views across 120 images, the errors in depth estimation introduced by side-view estimations, as well as the inaccuracies in the locations and shapes of eyes, nose, and mouth resulting from lateral-view estimations, accumulate and lead to substantial inaccuracies in both aspects. This phenomenon is

Figure 4: **Comparison with other 3D landmarks detection methods**. The first three columns display the 3D estimation results estimated by 3DDFA, averaged on all images, frontal images, and lateral images, respectively. The fourth to sixth columns show the 3D estimation results estimated by DECA, averaged on all images, frontal images, and lateral images, respectively. The last column presents our method's results.

Figure 5: **More visualization of accurate 3D landmarks detection of FLNeRF.**

observed in both the 2D estimation followed by triangulation and the averaged 3D estimation methods. Fig. 4 also demonstrates that our method outperforms other methods not only in estimating the locations and shapes of the mouth, eyes, nose, and cheeks but also in accurately predicting the depth of the facial profile.

Figure 6: More generalization examples on single in-the-wild images.

3 More Visualization Results

3.1 3D landmarks Detection On The NeRF

As extension of Figure 3 in our main paper, Fig. 5 shows more visualization results of accurate 3D landmarks prediction of our FLNeRF on face NeRFs. Observing the two figures, we can see how robust our FLNeRF is, that it predicts accurate 3D face landmarks for both males and females, people with various skin colors, faces under different illuminations, and even faces with glasses and beard.

3.2 Generalization On In-The-Wild Single Images

As stated in Section 3.5 in our main paper, our FLNeRF could be generalized to localize 3D face landmarks on face NeRFs reconstructed from a single in-the-wild face image leveraging EG3D Inversion (Chan et al., 2022). Fig. 6 shows more qualitative results as supplement to Figure 5 in our main paper. Since in-the-wild images vary from view directions, illumination, races, genders, make-up, and many complex and subtle factors, the accurate generalization results illustrate the power of our FLNeRF.

3.3 Video

By capitalizing our accurate 3D face landmarks, our modified MoFaNeRF could perform various downstream tasks, like face swapping and face editing introduced in section 4 in our main paper. Here we produce a video for more direct visualization. The video contains five parts:

1. **Accurate 3D face landmarks detection on NeRF.** Each row shows the visualization of the accurate 3D facial landmarks detection on the same identity from 3 different camera poses. The landmarks overlapped on the face NeRF are the estimated facial landmarks.
2. **Generalization on in-the-wild single images.** This part gives an intuitive illustration of how our FLNeRF could be generalized to detect accurate 3D landmarks on face NeRFs reconstructed from a single in-the-wild image as described in Section 3.5 in our main paper. In the video, we first show four in-the-wild images. Then we show reconstructed face NeRFs using EG3D inversion (Chan et al., 2022). Finally, we show overlayed 3D face landmarks predicted by our FLNeRF.
3. **Face editing by directly manipulating 3D face landmarks.** The two columns show the results obtained by manipulating 3D face landmarks on two different identities using our modified MoFaNeRF. The results are coherent in expression transitions and consistent in different view directions. The landmarks overlapped on the face NeRF are the target landmarks.
4. **3D face reenactment on NeRF.** For 3D face reenactment, we use FLNeRF to predict 3D face landmarks, given any face NeRF. The left face in the video is driver face NeRF with estimated landmarks overlaid. The predicted landmarks are fed together with the same person's texture map to our modified MoFaNeRF, which then produce the right face in the video.
5. **3D expression transfer on NeRF.** For 3D expression transfer, we use FLNeRF to predict 3D face landmarks on $\mathcal{I}_1$'s face NeRF. The left face in the video is driver face NeRF ($\mathcal{I}_1$'s

face NeRF) with estimated landmarks overlaid. The predicted landmarks are fed together with $\mathcal{I}_2$'s's texture map to our modified MoFaNeRF, which then produce the right face in the video ($\mathcal{I}_2$'s face NeRF).

This video demonstrates that our FLNeRF could produce accurate 3D face landmarks on NeRF. By leveraging EG3D Inversion (Chan et al., 2022), our FLNeRF could be well generalized to localize accurate 3D face landmarks on face NeRFs reconstructed from in-the-wild single images. Furthermore, with the help of our modified MoFaNeRF, FLNeRF could directly operate on dynamic NeRF, so an animator can easily edit, control, and even transfer emotion from another face NeRF.

4 TPS

The coefficients $\mathbf{A}_0$, $\mathbf{A}_1$ and ω_i mentioned in Section 4.2 can be found by solving the following linear system. Let $\mathbf{W} = [\omega_1, \cdots, \omega_N]$ and $\mathbf{Y} = [\mathbf{L'}^\intercal \ \mathbf{0} \ \mathbf{0} \ \mathbf{0}]^\intercal$:

$$\mathbf{K} = \begin{bmatrix} 0 & U(\|\mathbf{l}_1 - \mathbf{l}_2\|) & \cdots & U(\|\mathbf{l}_1 - \mathbf{l}_N\|) \\ U(\|\mathbf{l}_2 - \mathbf{l_1}\|) & 0 & \cdots & U(\|\mathbf{l}_2 - \mathbf{l}_N\|) \\ \cdots & \cdots & \cdots & \cdots \\ U(\|\mathbf{l}_N - \mathbf{l}_1\|) & U(\|\mathbf{l}_N - \mathbf{l}_2\|) & \cdots & 0 \end{bmatrix}_{N \times N} \tag{1}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{1} & \mathbf{l}_1^\intercal \\ \mathbf{1} & \mathbf{l}_2^\intercal \\ \cdots & \cdots \\ \mathbf{1} & \mathbf{l}_n^\intercal \end{bmatrix}_{N \times 4} \tag{2}$$

$$\mathbf{M} = \begin{bmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\intercal & \mathbf{0} \end{bmatrix}_{(N+4) \times (N+4)} \tag{3}$$

$$(\mathbf{W} | \mathbf{A}_1 \ \ \mathbf{A}_0)^\intercal = \mathbf{M}^{-1}\mathbf{Y} \tag{4}$$

5 More Ablation Studies of FLNeRF

Table 6: Since train/test data for coarse model only contains 20 basic expressions, we calculate the average Wing loss on these expressions for (a). For (b), (c) and (d), *whole face losses* are calculated on the test data set with 110 different expressions. *Mouth* and *Eyes* losses measure the corresponding landmarks' accuracy based on Wing loss. The last column shows results on basic mouth stretching expression and 10 augmented exaggerated expressions. All values are multiplied by 10.

	Average Wing Loss Using Bilinear Model				Average Wing Loss Using 3DMM			
	All Expressions			Exaggerated Expressions	All Expressions			Exaggerated Expressions
	Whole Face	Mouth	Eyes		Whole Face	Mouth	Eyes	
(a)	3.65±1.26	-	-	-	2.49±0.91	-	-	-
(b)	0.78±0.26	0.88±0.54	0.63±0.12	0.87±0.43	0.94±0.22	0.97±0.52	0.86±0.09	1.27±0.44
(c)	0.69±0.24	0.86±0.46	0.55±0.12	0.60±0.15	0.90±0.07	0.88±0.37	0.88±0.05	0.86±0.05
(d)	**0.63±0.20**	**0.77±0.43**	**0.55±0.13**	**0.53±0.12**	**0.86±0.08**	**0.87±0.39**	**0.85±0.06**	**0.84±0.08**

Table 6 tabulates ablation study results of our FLNeRF using VGG as backbone, while Table 2 in Section 4.3 in our main paper show statistics with VoxResNet as the backbone. We still conduct ablation on: (a) remove fine model, (b) remove expression augmentation, (c) use only two sampling scales, i.e., the first two rows in Figure 2 in the main paper, (d) our full model.

We follow the same test strategy as Section 4.3 to conduct this experiment. Similar to the results obtained by VoxResNet backbone, FLNeRF achieves the best among all ablation studies using VGG backbone.

6 Application Model

6.1 Model Architecture

Our modified MoFaNeRF model architecture is shown in Fig. 7, where we remove shape code, expression code, and ISM in the original MoFaNeRF model. This is because by (Zhuang et al.,

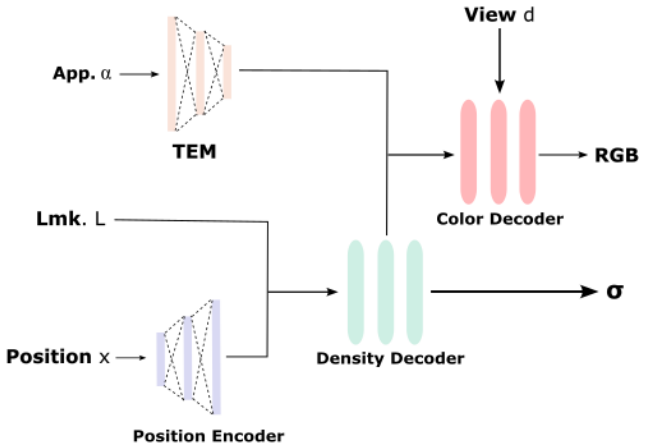

Figure 7: Architecture of **Modified MoFaNeRF model (2-code model)**, where we replace their shape and expression codes by our landmarks with position encoding. Appearance code which encodes the face texture remains the same. TEM is texture encoding module in (Zhuang et al., 2022).

2022)'s design, expression code is learnable, while shape code remains the same among all expressions of the same identity. However, face shape including location of mouth and eyebrows may also change during expression changes (e.g., brow raises, brow lowers, mouth twists to left or right, mouth stretches, jaw moves to left or right, etc). The combination of a static shape code with a learnable expression code may thus conflict with each other.

Instead, we directly concatenate 3D face landmarks to the encoded space position. The concatenated vector is fed into the density decoder. By doing so, our model takes in 3D space location, view direction, texture code, and 3D face landmarks as inputs to generate a face NeRF. Given texture map, we can render a face image with any given expression from any view points by manipulating the face landmarks. We believe that the original MoFaNeRF (Zhuang et al., 2022) attempts to extract deep information from each expression that is independent from the shape code, where such information mainly comes from 3D landmarks. That is why our application model outperforms MoFaNeRF a bit in terms of objective, structural, and perceptual similarity as validated in Table 4 in our main paper. Figure 6 in our main paper presents the qualitative results, showing that we can independently control movements of mouth, nose, eyes, and eyebrows by directly manipulating landmarks owing to our better disentanglement than (Zhuang et al., 2022) in their shape and expression codes.

Since our FLNeRF, which is trained on expanded data set with 110 expressions adopting the same training configurations as (Zhuang et al., 2022), can produce accurate 3D face landmark locations, and that our modified MoFaNeRF operates on landmarks directly, we can perform downstream tasks employing our face landmarks prediction on NeRF, i.e., face editing and face swapping.

6.2 Ablation Study

We perform ablation on: (a) original MoFaNeRF, (b) using three codes: texture, shape and landmarks; (c) our modified model which uses two codes: texture and landmarks. We render 300 images of the first 15 identities in our dataset (Zhu et al., 2021) for evaluation, where one image is synthesized for every expression and every identity with random view direction. Following (Zhuang et al., 2022), we use PSNR, SSIM and LPIPS criteria to assess objective, structural, and perceptual similarity. Table 7 tabulates the quantitative statistics on the corresponding coarse models.

Table 7: Quantitative evaluation of on our application model. (a) is original MoFaNeRF, (b) is our 3-code (texture, shape, landmark) MoFaNeRF model, (c) is our 2-code (texture, landmark) MoFaNeRF model. With our landmarks which effectively encode 3D shape, the original shape code in MoFaNeRf can be eliminated while outperforming (a) and (b). Values of SSIM and LPIPS are multiplied by 10.

	PSNR(dB)↑	SSIM↑	LPIPS↓	# params
(a)	24.85±1.91	8.53±0.35	1.72±0.32	29,100,936
(b)	21.71±1.46	7.28±0.50	3.50±0.42	29,584,776
(c)	**25.47±1.68**	**8.60±0.32**	**1.66±0.28**	29,456,776

From the testing statistics, our model outperforms (a) and (b). Interestingly, comparing (b) and (c), adding shape code as input substantially decreases performance, indicating the shape code does have redundant information with 3D landmarks. For a given identity with different expressions, the shape

Figure 8: Demonstration of **face editing** via direct landmark control. For each row, images are rendered by interpolating landmarks of the left most expression and the right most expression. This figure is an extension of Figure 8 in the main paper.

code remains the same while landmarks vary which confuses the network in (b). Comparing (a) with (c), 3D landmarks location alone outperform combination of shape and learnable expression code.

7 Ethics Discussion

Images we use for training, testing and visualization in this paper are from FaceScape (Zhu et al., 2021), an open-source dataset for research purpose. Our technology has the potential to cheat face recognition system. Therefore, it should not be abused for illegal purposes.

References

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022.

Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, Sep 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01521-4. URL http://dx.doi.org/10.1007/s11263-021-01521-4.

Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression, 2020.

Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *arXiv preprint arXiv:2111.01082*, 2021.

Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)*, 2022.