Portrait Neural Radiance Fields from a Single Image

Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang and Jia-Bin Huang

Virginia Tech, Google

Introduction

Portrait view synthesis

- camera poses using a light stage
- 3D model based methods only covers the center of the face

NeRF

- modeling the volumetric density and color by MLP
- requires images of static subjects from multiple viewpoints

Introduction

Propose

- Train an MLP for a single headshot portrait.

Pretrain MLP

- a na¨ıve pretraining performs poorly for unseen subjects

Meta learning

- adapts to an unseen subject

Rigid transform

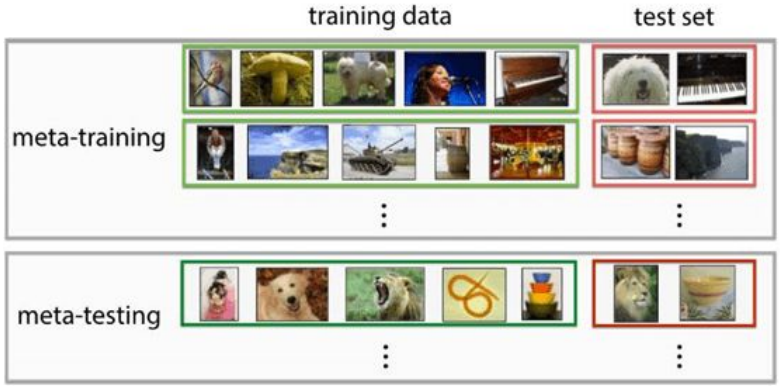
- using a rigid transform from the world coordinate in a canonical face space

Dataset

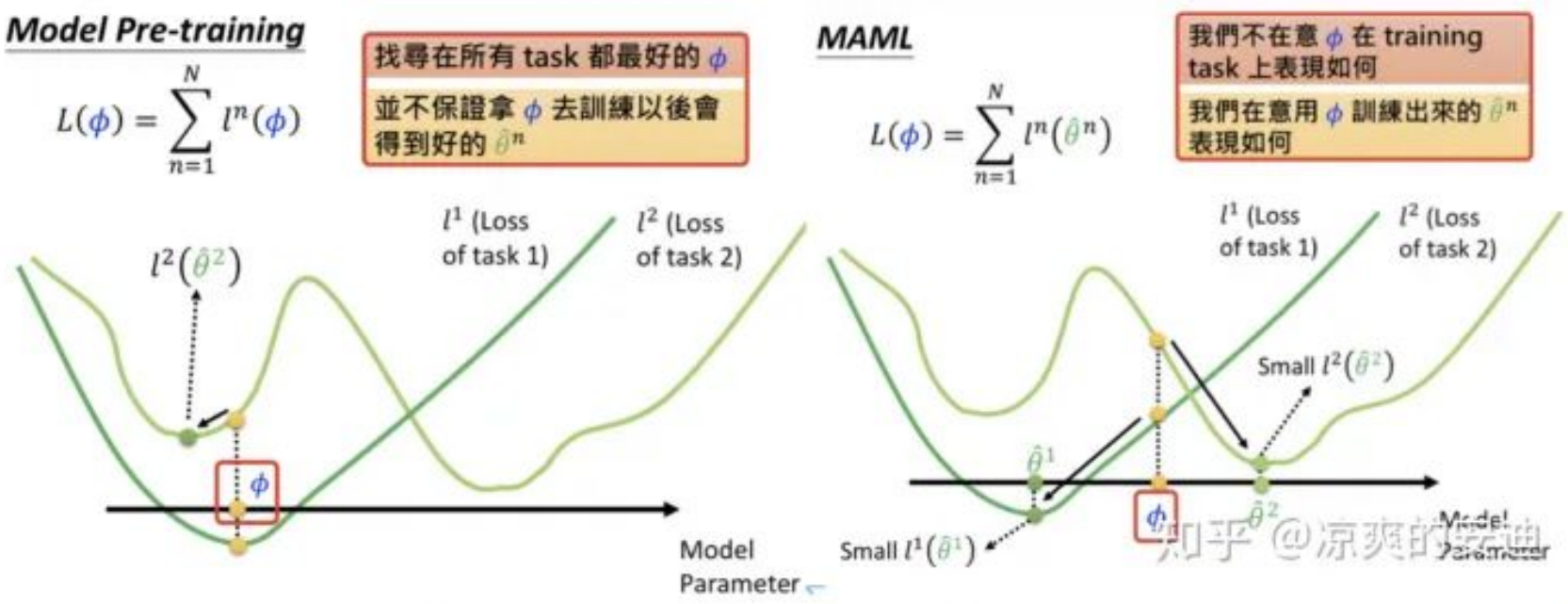
- multi-view protrait dataset in a light stage

Meta-Learning

- learn-to-learn
- Machine Learning & Meta Learning
  - ML : y = f(x)
  - Meta-Learning : f = F*, y= f(x)
- Support set (training set), Query set (testing set)
- Task : Training task & Testing task



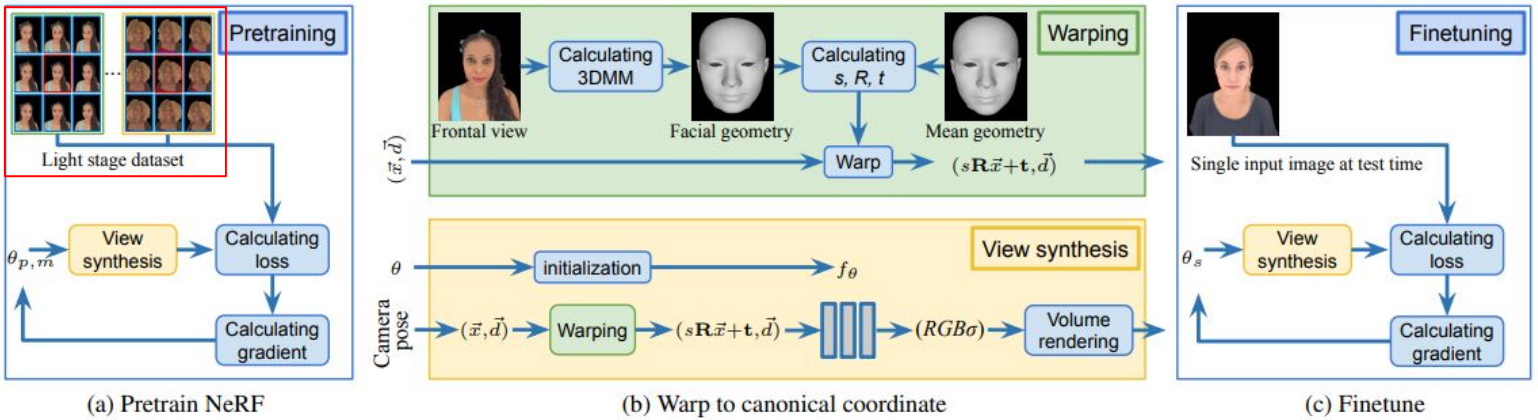
Meta-Learning



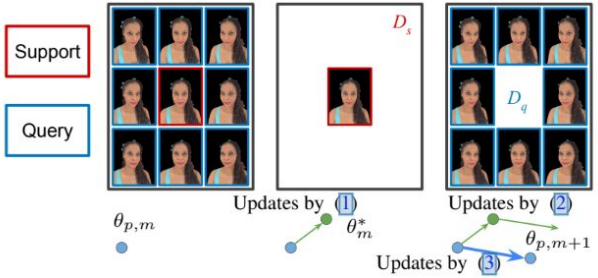
Algorithm

Training data

- light stage captures over multiple subjects
- Support set (D_s) : The center view to the front view expected at the test time.
- Query set (D_q) : The remaining views are the target for view synthesis.
- Task (T_m) : NeRF model parameter for subject m from the D_s



Algorithm



Pretraing NeRF

- Goal : Pretrain a NeRF model parameter θ_p^*.
- Loop K subjects, model parameter in each subject m as $\theta_{p,m}, m = \{0, ...K-1\}$
- For each task, train D_s and D_q

Pretrain D_s

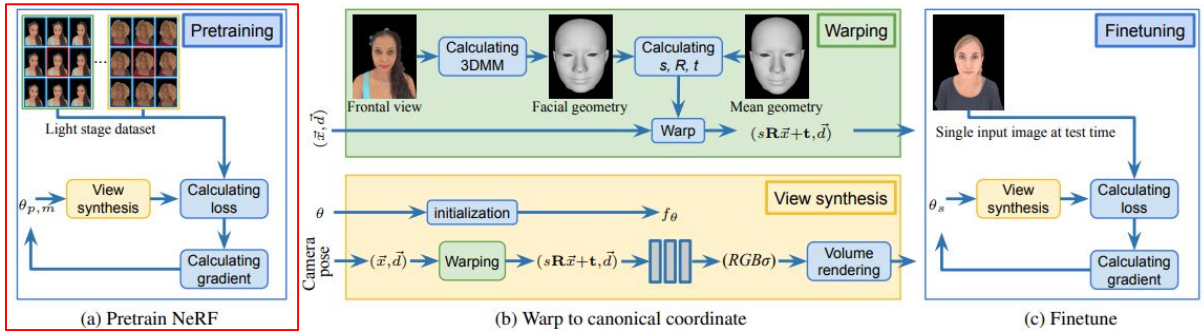
$$\theta_m^{t+1} = \theta_m^t - \alpha \nabla_\theta L_{D_s}(f_{\theta_m^t})$$

Pretrain D_q

$$\theta_m^{t+1} = \theta_m^t - \beta \nabla_\theta L_{D_q}(f_{\theta_m^t})$$

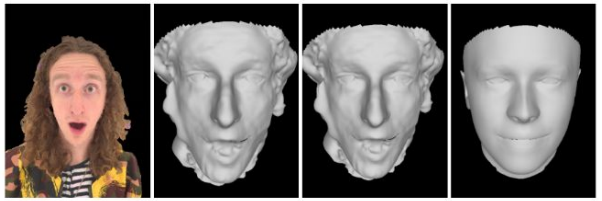
$$\theta_{p,m}^{t+1} = \theta_{p,m}^t - \beta \nabla_\theta L_{D_q}(f_{\theta_m^t})$$

$$\theta_{p,m} = \theta_{p,m}^{N_q - 1}$$

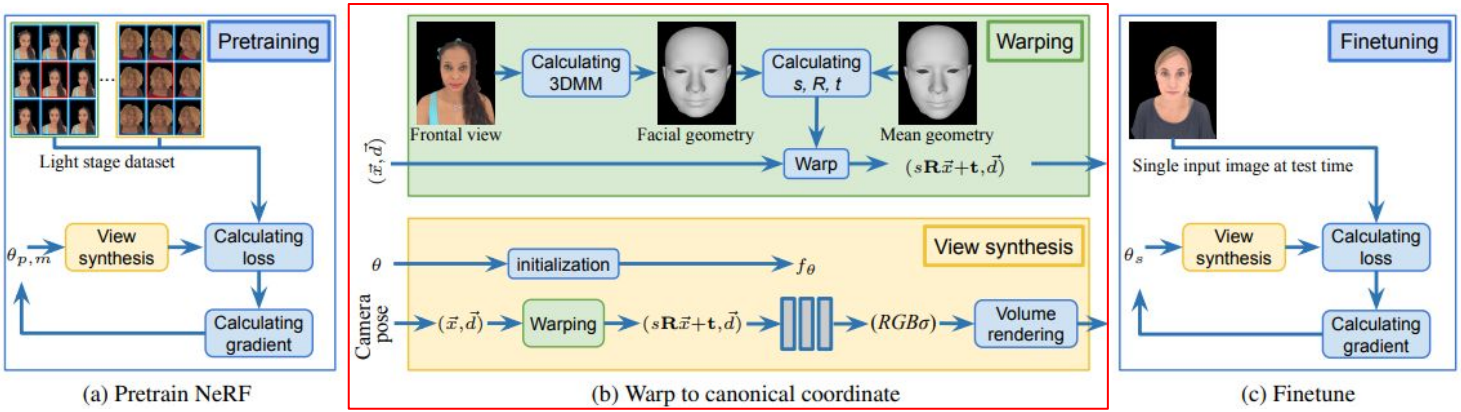


Algorithm

Canonical face space



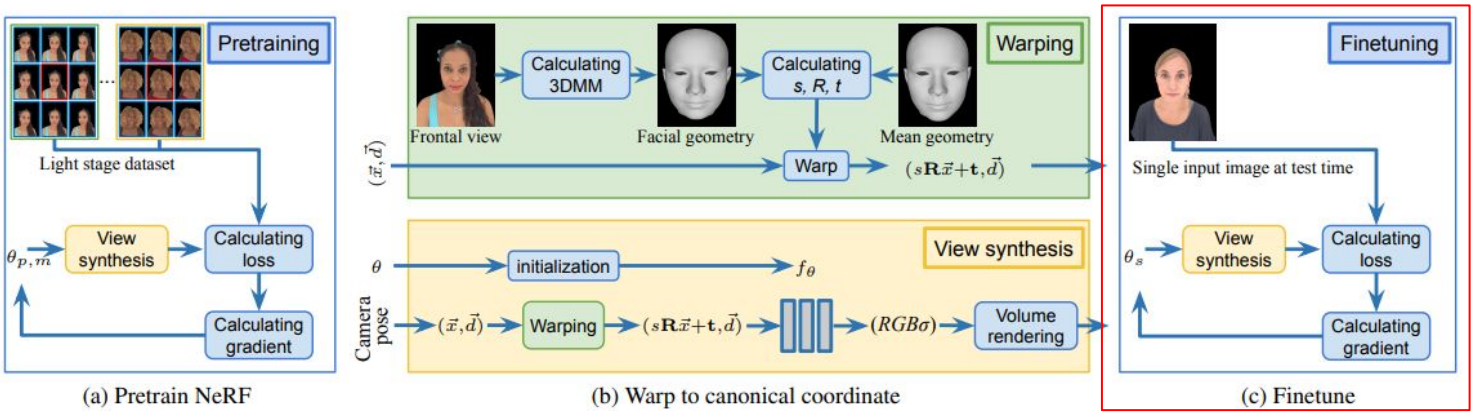
- normalize the world coordinate to canonical space
- rigid transform : $x' = s_m R_m x + t_m$
- use SVD decomposition to optimize rigid transform between F_m and $\bar{F}$



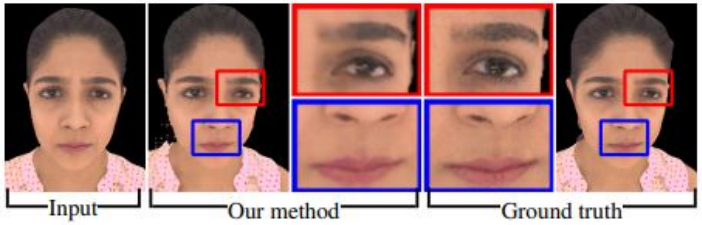
Algorithm

Finetuning and rendering

- a single frontal view of the subject s
- rigid transform between the world and canonical coordinate.
- finetune the pretrained model parameter
- sample the camera ray in the 3D space, warp to the canonical space



Experimental Results



Experimental Results

PSNR :

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

$$PSNR = 10 \cdot log_{10}(\frac{MAX_I^2}{MSE})$$

SSIM :

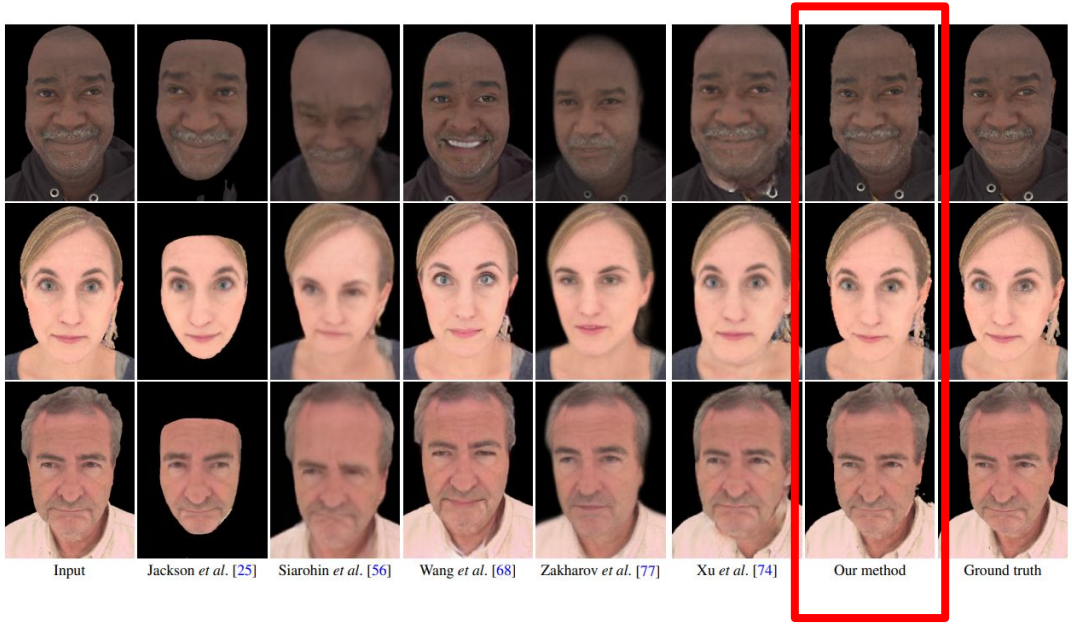
$$l(x,y) = \frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1} c(x,y) = \frac{2\sigma_x\sigma_y+c_2}{\sigma_x^2+\sigma_y^2+c_2} s(x,y) = \frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3}$$

$$SSIM(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma]$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$$

LPIPS

- use pre-defined network to computers the image similarity

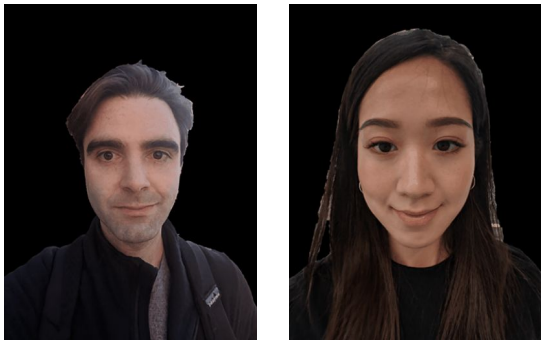


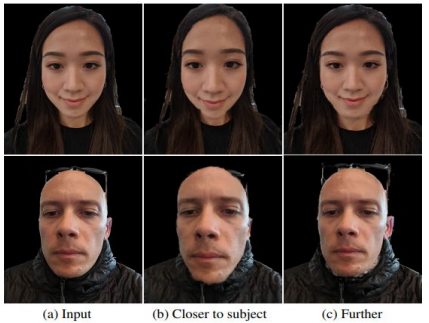
	PSNR ↑	SSIM ↑	LPIPS ↓
Jackson *et al.* [25]	10.23	0.4356	0.485
Wang *et al.* [68]	14.70	0.4578	0.380
Zakharov *et al.* [77]	15.25	0.4746	0.403
Siarohin *et al.* [56]	15.90	0.5149	0.437
Xu *et al.* [74]	18.91	0.5609	0.276
Our method	**23.92**	**0.7688**	**0.161**

Experimental Results

Perspective manipulation

- moving the camera from the subject and adjusting the focal length





(a) Input (b) Closer to subject (c) Further

Initialization

- compares the results from different initialization methods.

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_m \mathcal{L}_{\mathcal{D}_s}(f_\theta) + \mathcal{L}_{\mathcal{D}_q}(f_\theta),$$

(a) Random (b) Pretrain by (5) (c) Our method (d) GT

Table 2. **Ablation study on initialization methods.**

Initialization	PSNR ↑	SSIM ↑	LPIPS ↓
Random	14.99	0.5763	0.491
Pretrain by (5)	22.87	0.7824	0.215
Our method	**23.70**	**0.8051**	**0.178**

Experimental Results

Canonical face coordinate

(a) Input (b) World coordinate (c) Our method

Table 3. **Ablation study on canonical face coordinate.**

Coordinate	PSNR ↑	SSIM ↑	LPIPS ↓
World	24.80	0.8167	0.172
Canonical (our method)	**24.98**	**0.8178**	**0.156**

Input views in test time

(a) 1 view PSNR = 24.98 (b) 2 views PSNR = 27.70 (c) 5 views PSNR = 32.45 (d) GT

Limitaiton

(a) Background (b) Non-frontal view