CSC2457 3D & Geometric Deep Learning

Deformable Neural Radiance Fields

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz,
Dan B Goldman, Steven M. Seitz, Ricardo Martin-Brualla

Date: March 2, 2021

Presenter: Yun-Chun Chen

Instructor: Animesh Garg

UNIVERSITY OF TORONTO

Motivation and Main Problem

(a) Capture Process (b) Input (c) Nerfie (d) Nerfie Depth

Photorealistically reconstructing a non-rigidly deforming scene using photos/videos captured casually from mobile photos

Motivation and Main Problem

Applications:

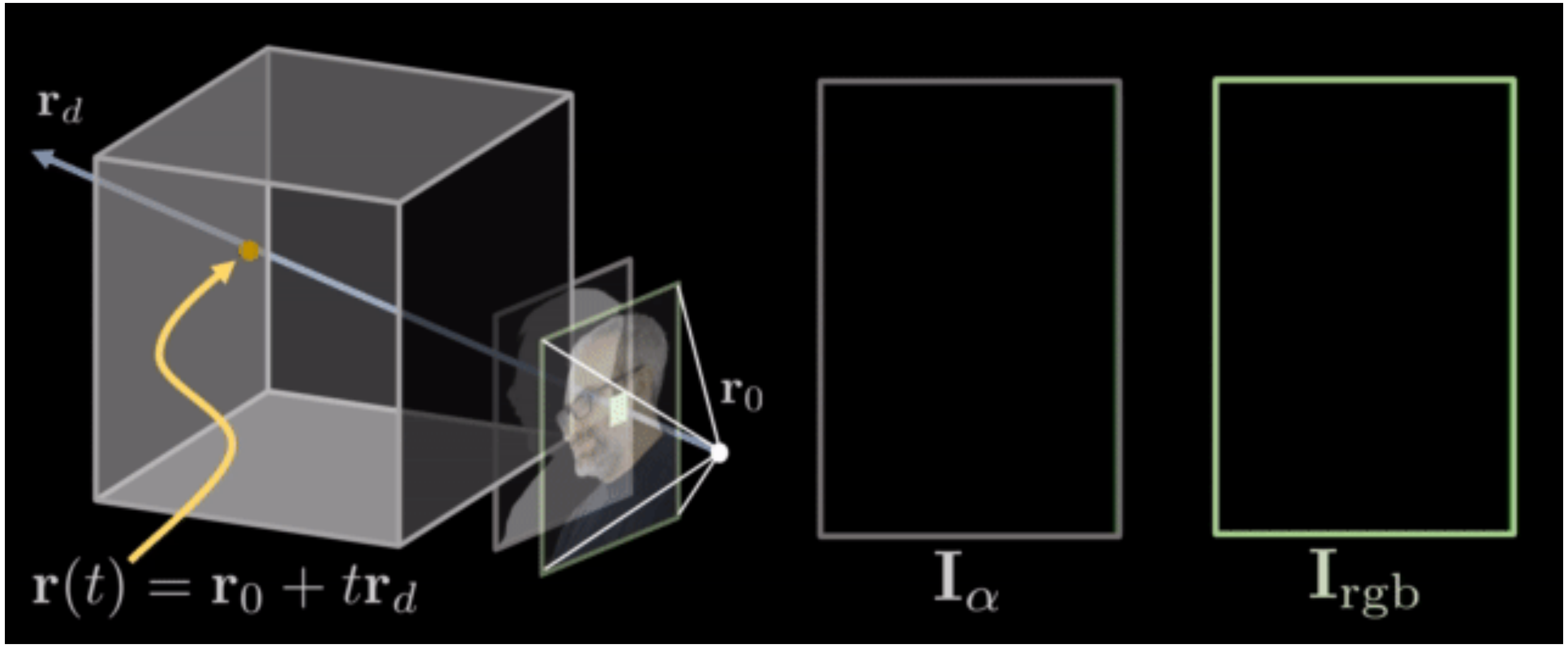
- Increased accessibility and applications of 3D modeling technology

Challenges:

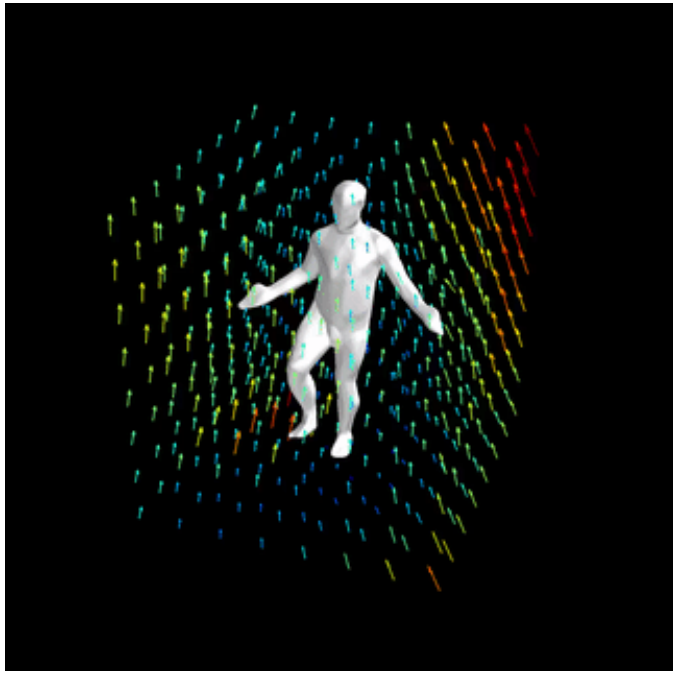
- Nonrigidity: our inability to stay perfectly still
- Challenging materials like hair, glasses, and earrings

Prior Work: Non-rigid Reconstruction

Neural Volumes

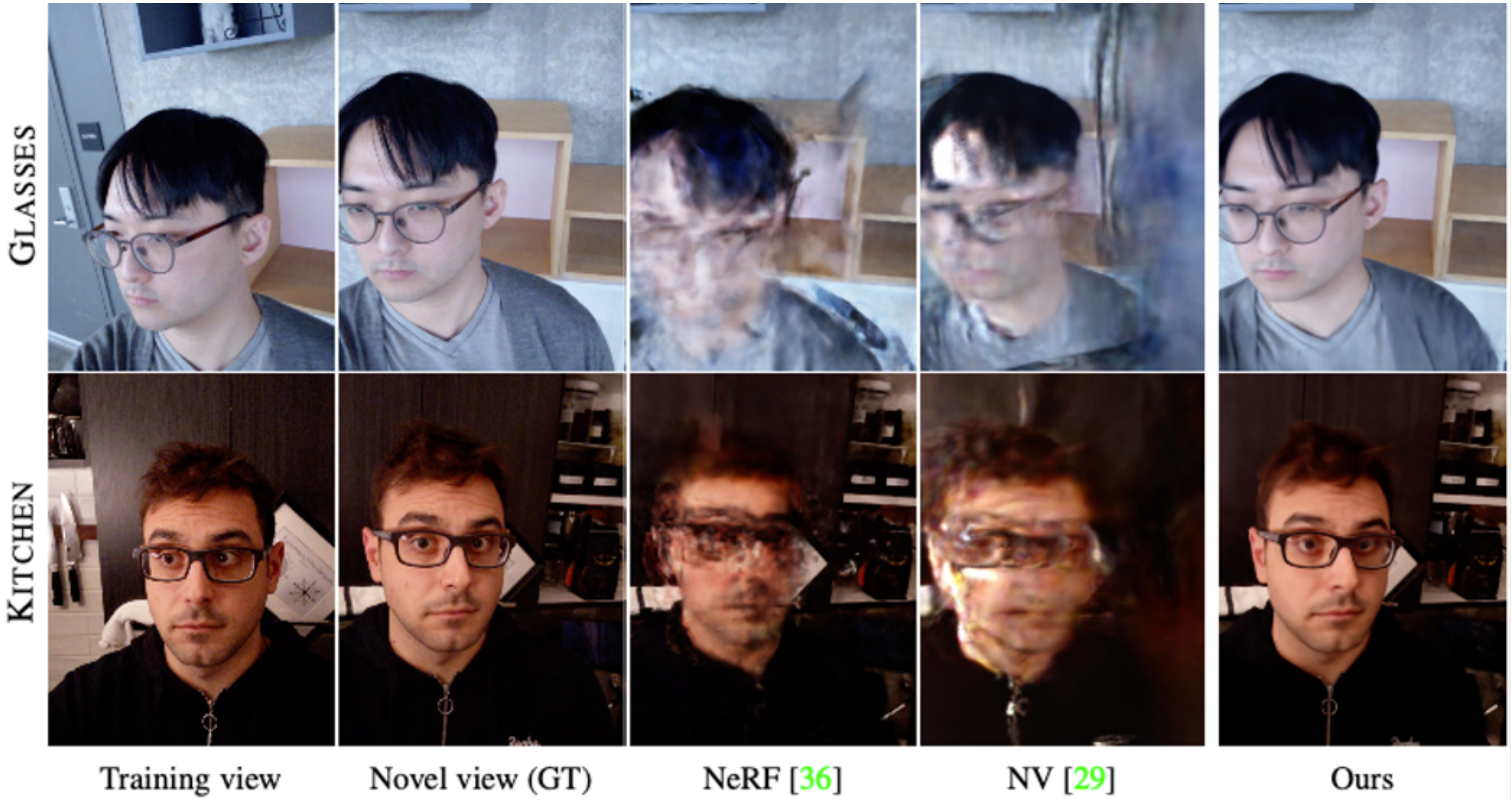


OccFlow



Limitations of Prior Work

- Cannot handle non-rigidly deforming scenes



Contributions

- A method for generating photorealistic novel views of humans

- A canonical NeRF model as a template for all observations

- A deformation field for 3D point warping

- High-fidelity reconstructions

General Background

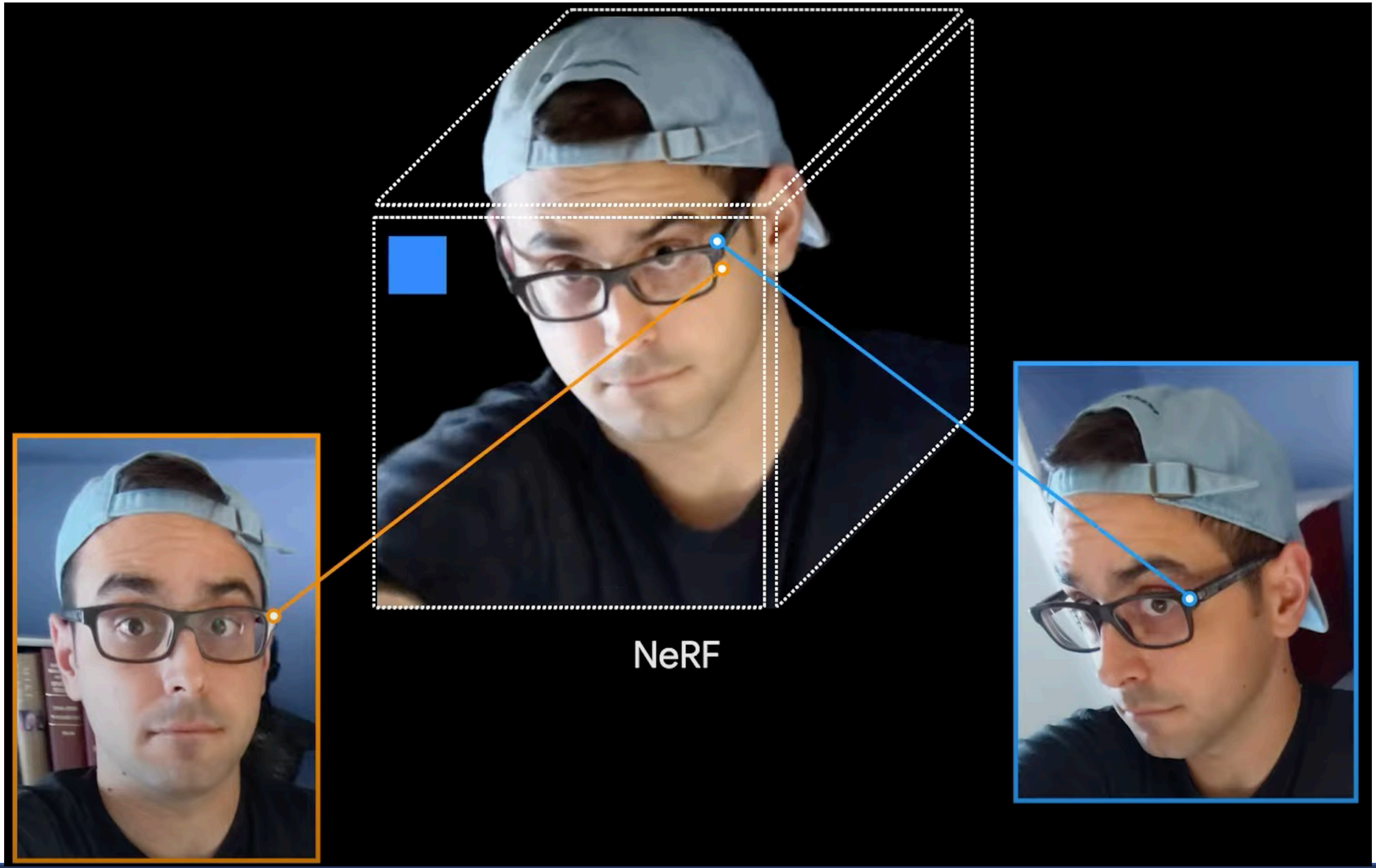
NeRF: $F: (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$

NeRF-A: $F: (\mathbf{x}, \mathbf{d}, \psi_i) \rightarrow (\mathbf{c}, \sigma)$

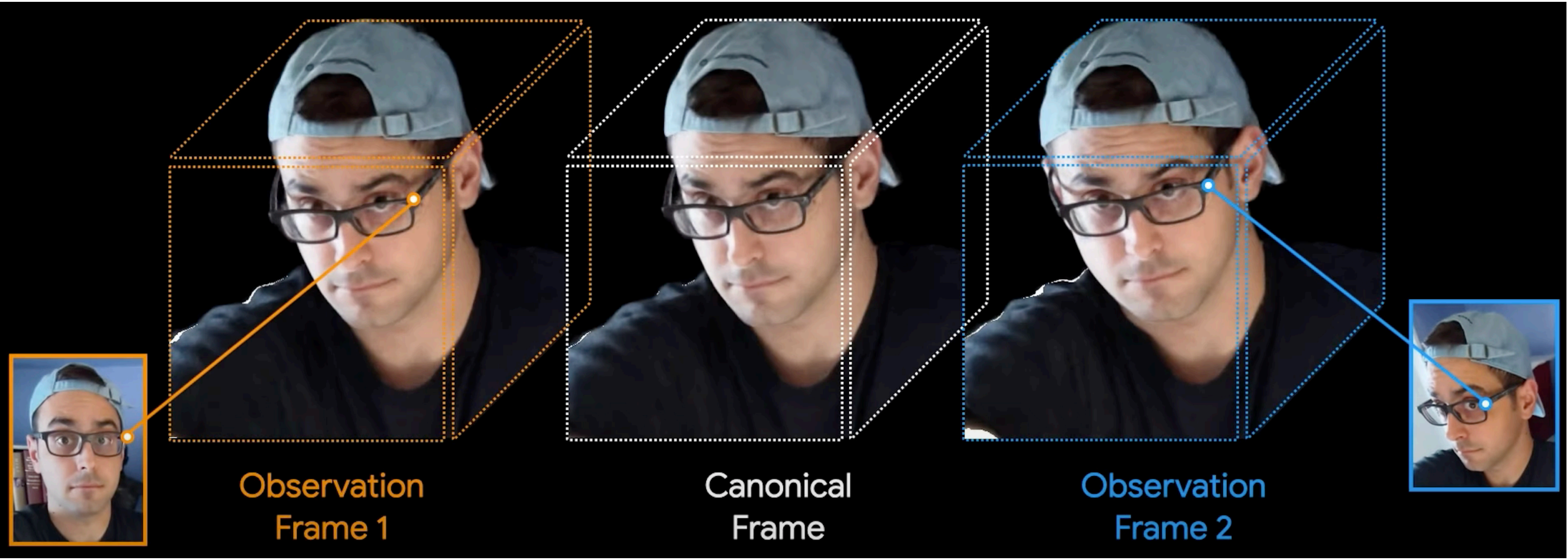
Notations:

- $\mathbf{x}$: 3D position
- $\mathbf{d}$: viewing angle
- $\mathbf{c}$: color
- σ: density
- ψ_i: appearance code for each observed frame i

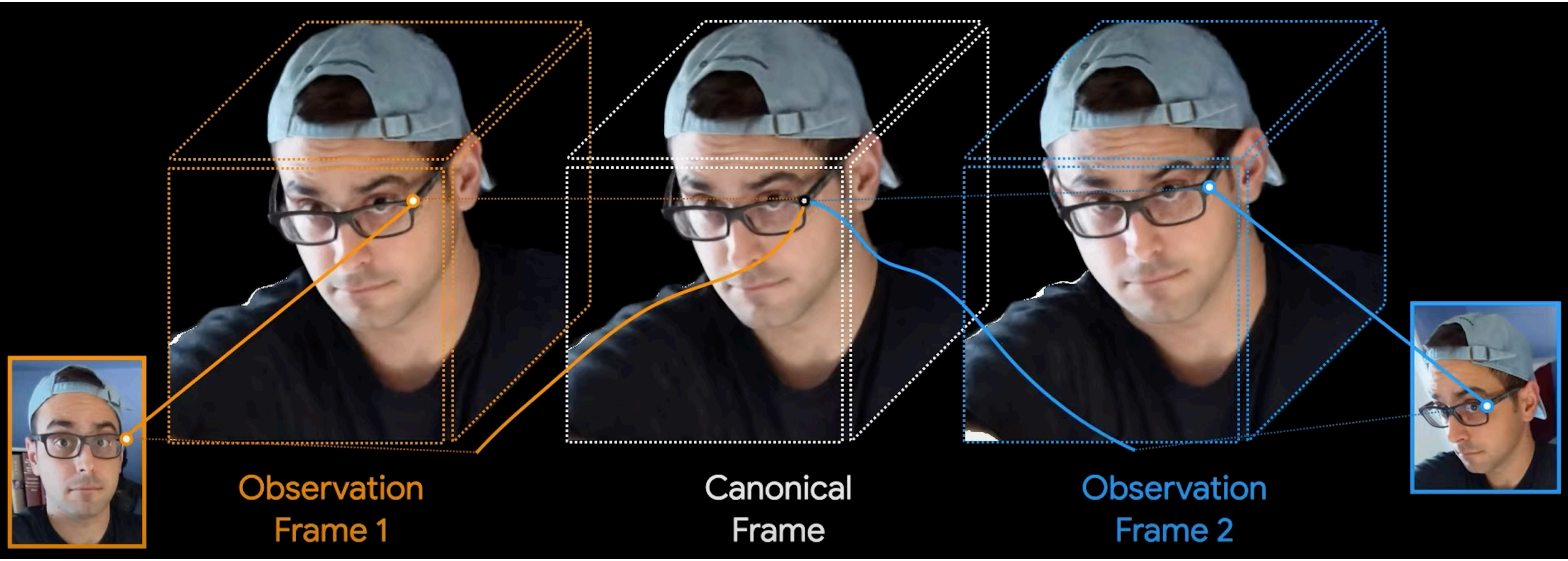
Motivation and Observation



Different Observation Frames



Canonical Frame



Problem Setting

NeRF: $F: (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$

NeRF-A: $F: (\mathbf{x}, \mathbf{d}, \psi_i) \rightarrow (\mathbf{c}, \sigma)$

Notations:

- $\mathbf{x}$: 3D position
- $\mathbf{d}$: viewing angle
- $\mathbf{c}$: color
- σ: density
- ψ_i: appearance code for each observed frame i

Problem Setting

NeRF: $F: (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$

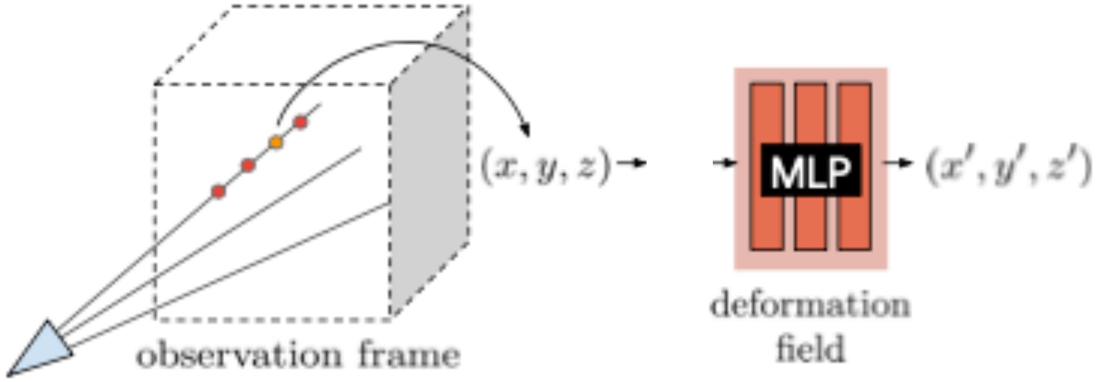
NeRF-A: $F: (\boxed{\mathbf{x}}, \mathbf{d}, \psi_i) \rightarrow (\mathbf{c}, \sigma)$

D-NeRF: $F: (\boxed{T(\mathbf{x}, \omega_{\boldsymbol{i}})}, \mathbf{d}, \psi_i)$

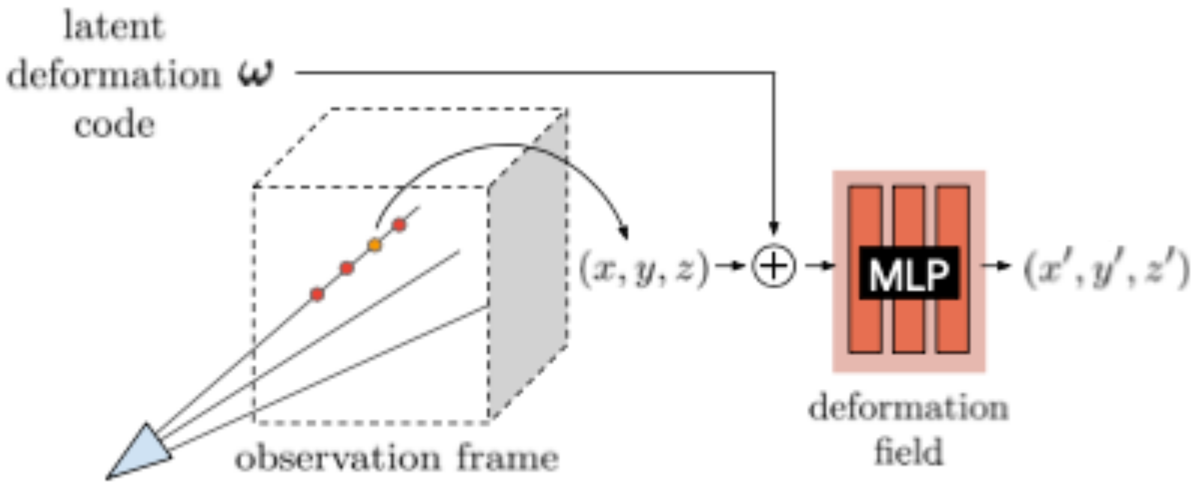
Notations:

- $\mathbf{x}$: 3D position
- $\mathbf{d}$: viewing angle
- $\mathbf{c}$: color
- σ: density
- ψ_i: appearance code for each observed frame i
- T: observation-to-canonical mapping
- $\omega_{\boldsymbol{i}}$: per-frame learned latent code

Approach



Approach



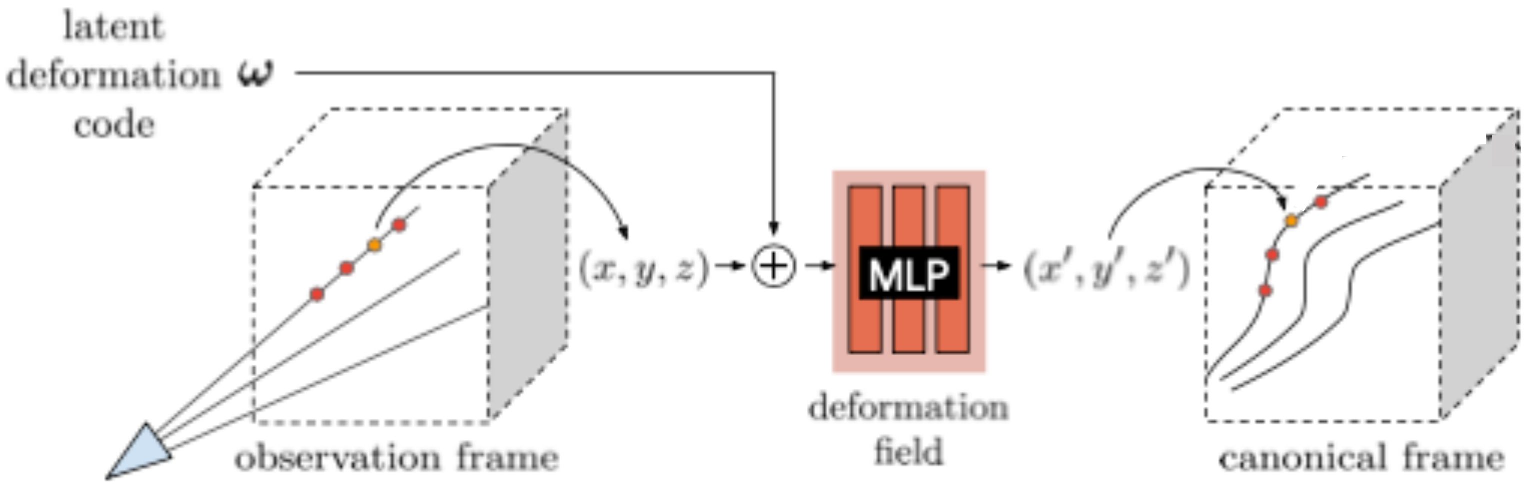
- Deformation field: SE(3)
- SE(3) transform: rotation **q** with pivot point **s** followed by a translation **t**

$$\mathbf{q} = \exp(\mathbf{p}) = \begin{pmatrix} \cos\|\mathbf{v}\| \\ \frac{\mathbf{v}}{\|\mathbf{v}\|}\sin\|\mathbf{v}\| \end{pmatrix}$$

$$\mathbf{x}' = \mathbf{q}(\mathbf{x} - \mathbf{s})\mathbf{q}^{-1} + \mathbf{s} + \mathbf{t}$$

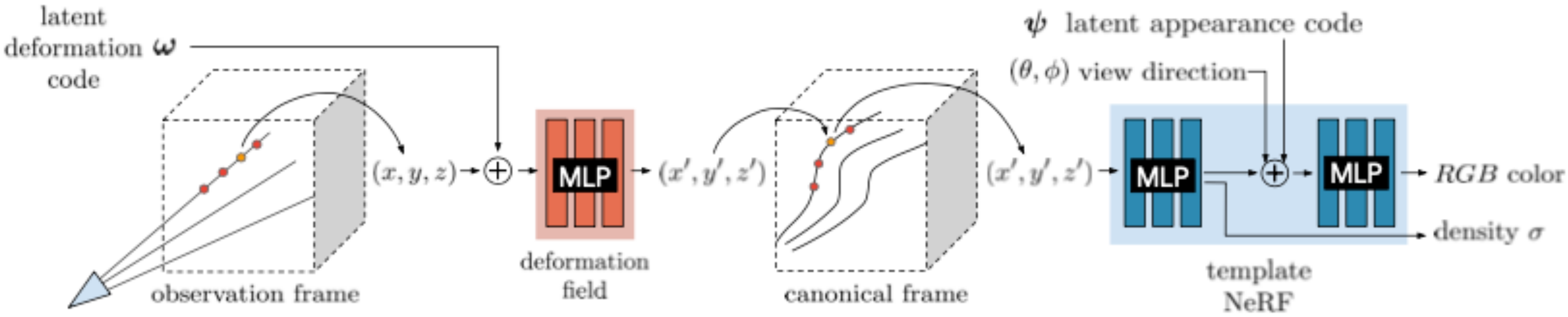
- MLP: $(\mathbf{x}, \boldsymbol{\omega}_i) \rightarrow (\mathbf{v}, \mathbf{s}, \mathbf{t})$

Approach



latent
deformation $\boldsymbol{\omega}$
code
(x, y, z)
MLP
(x', y', z')
deformation
field
observation frame
canonical frame

Approach



latent deformation code ω
observation frame
(x, y, z)
MLP
deformation field
(x', y', z')
canonical frame
(x', y', z')
ψ latent appearance code
(θ, ϕ) view direction
MLP
MLP
RGB color
density σ
template NeRF

Elastic Regularization

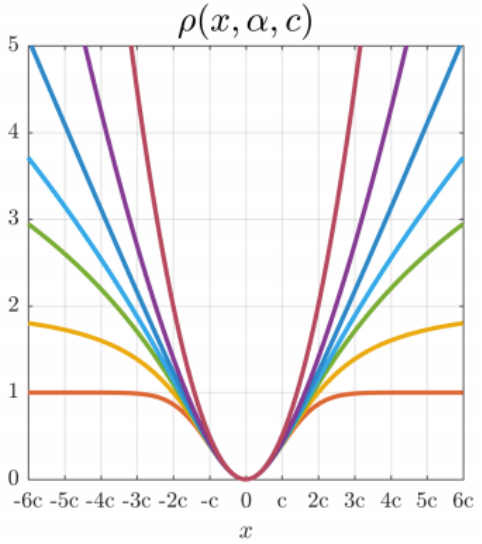
- The deformation field adds ambiguities
- Solution: use elastic energies
- Goal: achieve local rigidity

Elastic Regularization

- Compute the Jacobian for each point $\mathbf{J}_T(\mathbf{x})$
- Apply SVD: $\mathbf{J}_T(\mathbf{x}) = \boldsymbol{U\Sigma V^T}$
- Measure the deviation of the singular values of $\mathbf{J}_T$ from the identity

$$L_{\text{elastic}}(\mathbf{x}) = \|\log \boldsymbol{\Sigma} - \log \mathbf{I}\|_F^2 = \|\log \boldsymbol{\Sigma}\|_F^2$$

Elastic Regularization



- Compute the Jacobian for each point $\mathbf{J}_T(\mathbf{x})$
- Apply SVD: $\mathbf{J}_T(\mathbf{x}) = \boldsymbol{U\Sigma V^T}$
- Measure the deviation of the singular values of $\mathbf{J}_T$ from the identity

$$L_{\text{elastic}}(\mathbf{x}) = \|\log \mathbf{\Sigma} - \log \mathbf{I}\|_F^2 = \|\log \mathbf{\Sigma}\|_F^2$$

- Robustness: remap the elastic energy defined above with a robust loss

$$L_{\text{elastic-r}}(\mathbf{x}) = \rho\left(\|\log \mathbf{\Sigma}\|_F, c\right),$$

$$\rho(x, c) = \frac{2(x/c)^2}{(x/c)^2 + 4}.$$

Background Regularization

- The deformation field is unconstrained
- Add a regularization term to prevent the background from moving

$$L_{\mathrm{bg}} = \frac{1}{K}\sum_{k=1}^{K}\left\|T(\mathbf{x}_k) - \mathbf{x}_k\right\|_2$$

Coarse-to-Fine Deformation Regularization

- Positional encoding: $\mathbb{R}^3 \to \mathbb{R}^{3+6m}$

$$\gamma(\mathbf{x}) = \left(\mathbf{x}, \cdots, \sin\left(2^k \pi \mathbf{x}\right), \cos\left(2^k \pi \mathbf{x}\right), \cdots\right)$$

- Higher m: higher frequency details, but may result in overfitting and modeling image noise
- Smaller m: not able to model deformations which require high frequency details

Coarse-to-Fine Deformation Regularization

- Positional encoding: $\mathbb{R}^3 \to \mathbb{R}^{3+6m}$

$$\gamma(\mathbf{x}) = \left(\mathbf{x}, \cdots, \sin\left(2^k \pi \mathbf{x}\right), \cos\left(2^k \pi \mathbf{x}\right), \cdots\right)$$

- Coarse to fine:

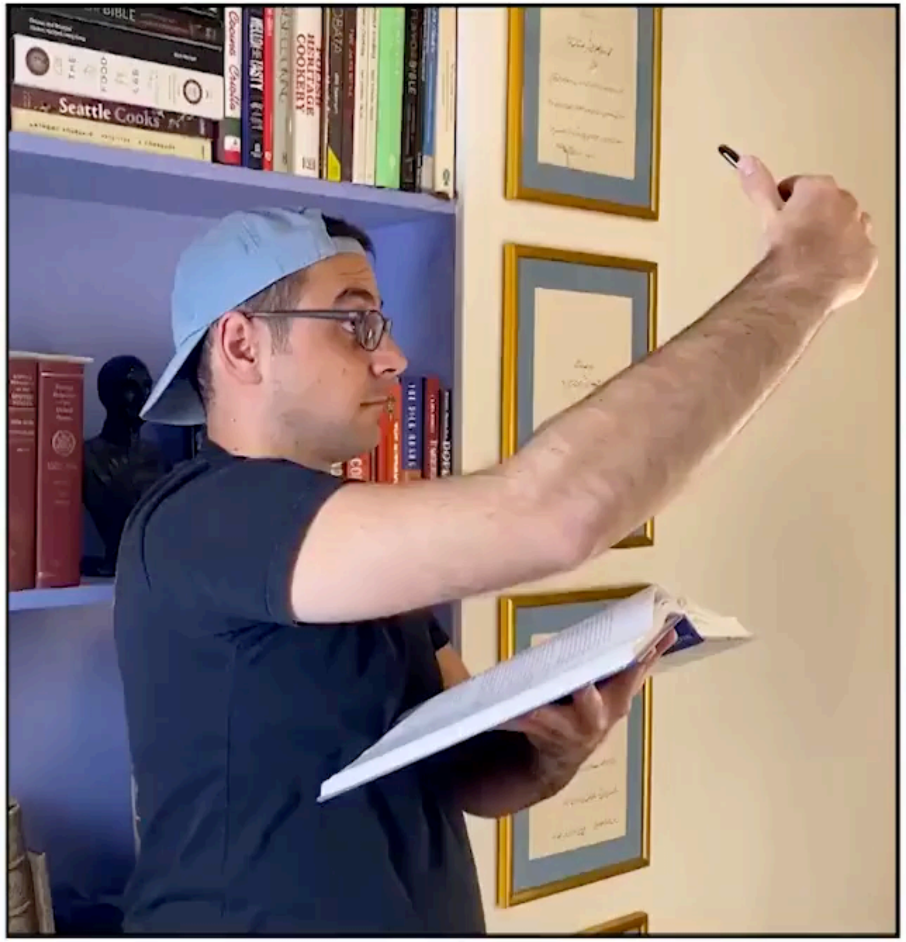
$$\gamma_\alpha(\mathbf{x}) = \left(\mathbf{x}, \cdots, \boxed{w_k(\alpha)} \sin\left(2^k \pi \mathbf{x}\right), \boxed{w_k(\alpha)} \cos\left(2^k \pi \mathbf{x}\right), \cdots\right)$$

$$w_j(\alpha) = \frac{(1 - \cos(\pi \operatorname{clamp}(\alpha - j, 0, 1))}{2} \qquad \alpha(t) = \frac{mt}{N}$$

Nerfies: Casual Free-Viewpoint Selfies

- Application: reconstruct high quality models of humans from casually captured selfies

- Input: a sequence of selfie photos or a selfie video (user is standing mostly still)



Nerfies: Casual Free-Viewpoint Selfies

- Frame selection: filter blurry frames using the variance of the Laplacian

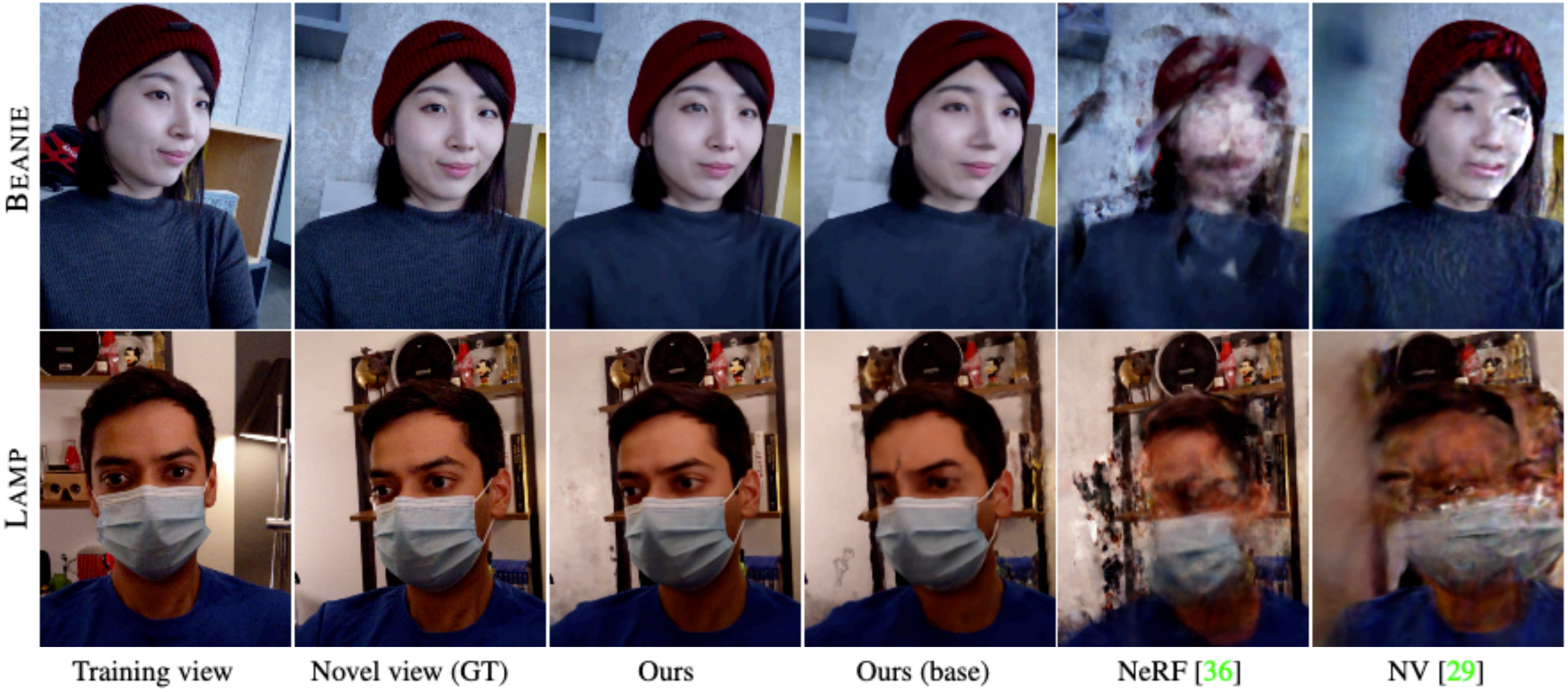
- Camera registration: use SfM to compute camera poses for each image and intrinsic calibration

- Foreground segmentation: use a foreground segmentation network to filter out features on the subject

Experimental Results

	Glasses (78 images)			Beanie (74 images)			Curls (57 images)			Kitchen (40 images)			Lamp (55 images)			Mean		
	PSNR↑	MS-SSIM↑	LPIPS↓	PSNR↑	MS-SSIM↑	LPIPS↓	PSNR↑	MS-SSIM↑	LPIPS↓	PSNR↑	MS-SSIM↑	LPIPS↓	PSNR↑	MS-SSIM↑	LPIPS↓	PSNR↑	MS-SSIM↑	LPIPS↓
NeRF [36]	17.69	.5962	.4723	16.58	.5524	.5884	14.28	.4517	.5921	18.79	.6873	.4094	17.42	.6447	.4268	16.95	.5865	.4978
NeRF + latent	21.76	.8201	.3239	20.89	.7711	.4235	22.20	.8040	.3446	21.24	.8212	.3075	20.63	.8489	.2364	21.34	.8131	.3272
Neural Volumes [29]	15.62	.5217	.5759	15.82	.5807	.5630	15.26	.5421	.5506	14.84	.5533	.5719	13.56	.5194	.5558	15.02	.5434	.5635
Ours	24.78	.8783	.2354	23.04	.8338	.3444	24.08	.8613	.2526	23.48	.8759	.2299	22.08	.8729	.1807	23.49	.8644	.2486
No elastic	24.61	.8760	.2357	23.22	.8356	.3451	23.75	.8527	.2547	23.28	.8729	.2393	21.96	.8726	.1801	23.36	.8620	.2510
No coarse-to-fine	23.51	.8434	.2551	21.41	.7875	.3684	23.08	.8284	.2939	23.11	.8667	.2455	22.51	.8751	.1876	22.72	.8402	.2701
No background reg.	24.20	.8656	.2360	19.47	.6989	.3904	20.73	.7620	.2964	21.83	.8395	.2569	19.82	.8078	.2061	21.21	.7947	.2772
Ours (base)	23.91	.8479	.2711	21.83	.7816	.4046	22.85	.8224	.3069	22.21	.8209	.3049	21.92	.8571	.2202	22.54	.8260	.3015

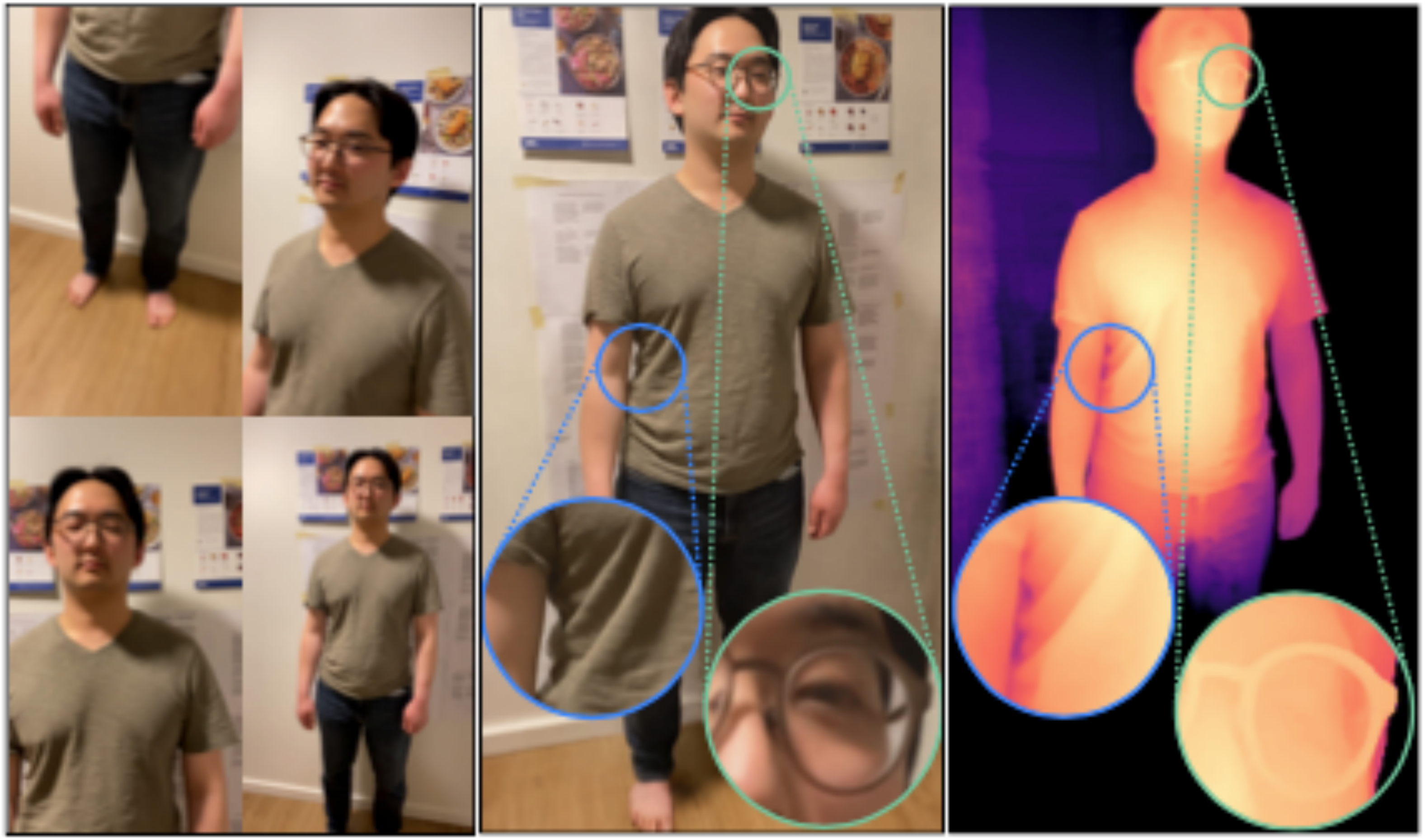
Experimental Results



Experimental Results

Input Video | Novel View Color | Novel View Depth

Experimental Results



example inputs rendered rgb rendered depth

Experimental Results

input images ground truth rendered color rendered depth

Discussion of results

- Renders novel views of humans with photorealistic quality

- Details (e.g., hair) are recovered

- Outperforms NeRF and NV

- Does not rely on domain specific priors (e.g., the dog example)

Critique / Limitations / Open Issues

- Can the method handle larger deformations that include full body motions?

- What would happen if the captured data is under lighting variations?

- What if background is also moving?

- How much data is needed (density of capture)?

Contributions (Recap)

- A method for generating photorealistic novel views of humans

- A canonical NeRF model as a template for all observations

- A deformation field for 3D point warping

- High-fidelity reconstructions