Summer Seminar 2022

그림쟁이들은 그림자까지 그려

김기훈

Table of Contents

After this seminar..

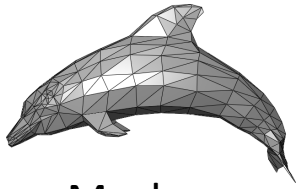
- Neural rendering

- Application: Neural rendering + Human body/face models

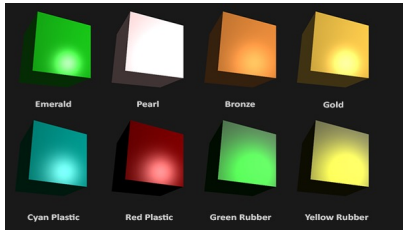
Neural Rendering

- Computer Graphics + Machine Learning
  - Generate photo-realistic imagery in a controllable way

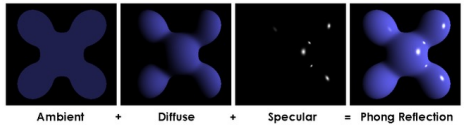
- Computer Graphics



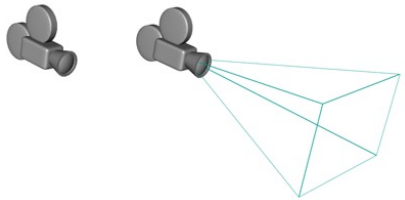
Meshes



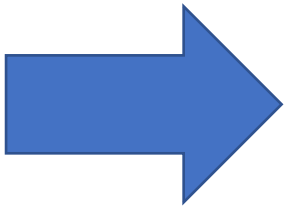
Material



Phong Shading



Camera parameters

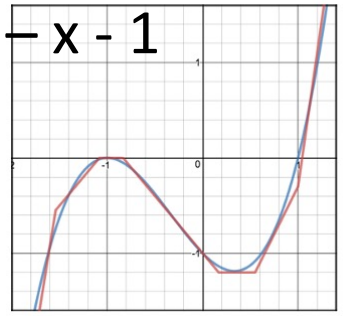


Real world scene

Neural Rendering

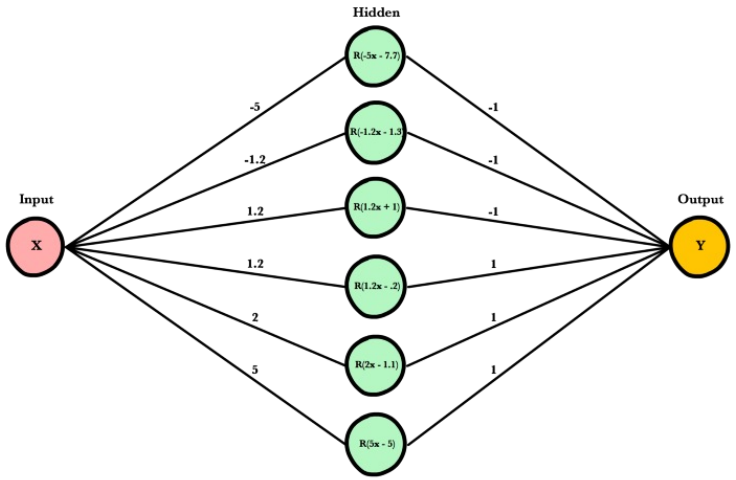
- Machine Learning
  - Universal function approximator

$y = x^3 + x^2 - x - 1$

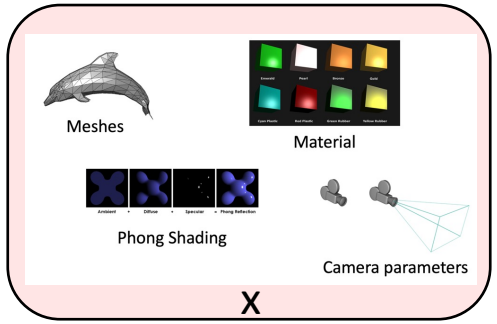


$$n_1(x) = Relu(-5x - 7.7)$$
$$n_2(x) = Relu(-1.2x - 1.3)$$
$$n_3(x) = Relu(1.2x + 1)$$
$$n_4(x) = Relu(1.2x - .2)$$
$$n_5(x) = Relu(2x - 1.1)$$
$$n_6(x) = Relu(5x - 5)$$

$$Z(x) = -n_1(x) - n_2(x) - n_3(x) + n_4(x) + n_5(x) + n_6(x)$$



Sparse data interpolation problem



ML

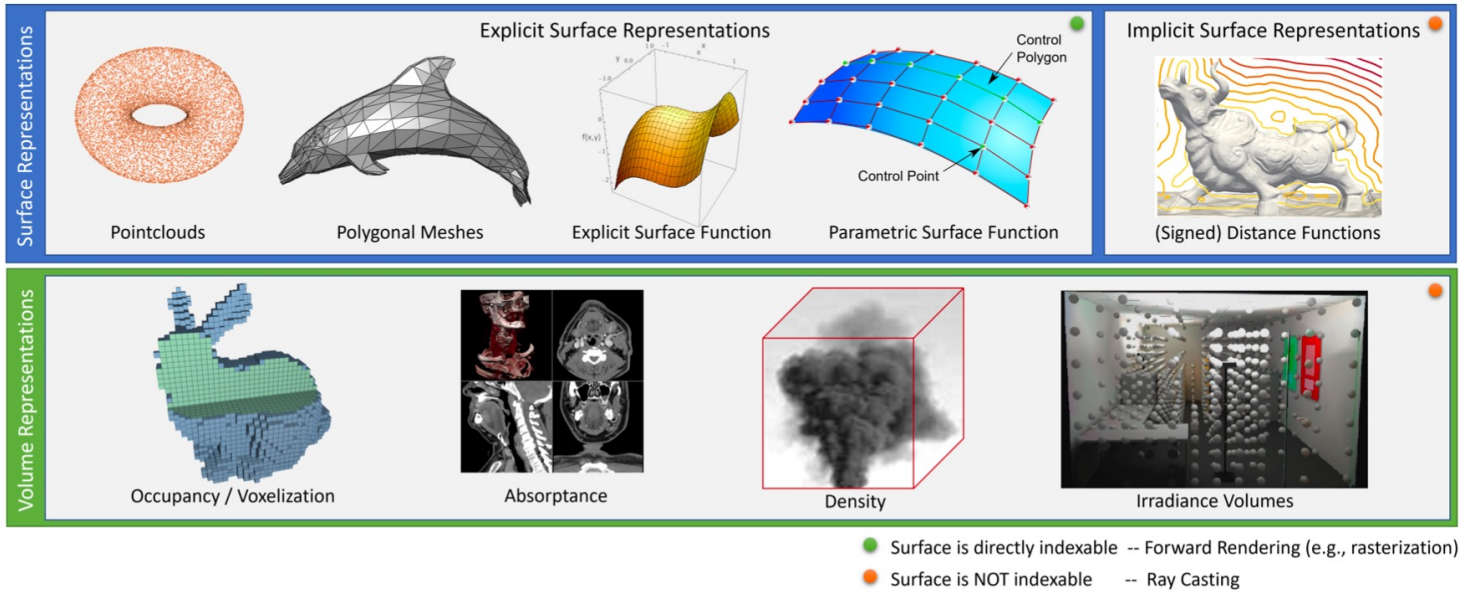
Neural Rendering

- Scene representation

$$S_{explicit} = \left\{ \begin{pmatrix} x \\ y \\ f_{explicit}(x,y) \end{pmatrix} \middle| \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \right\}$$

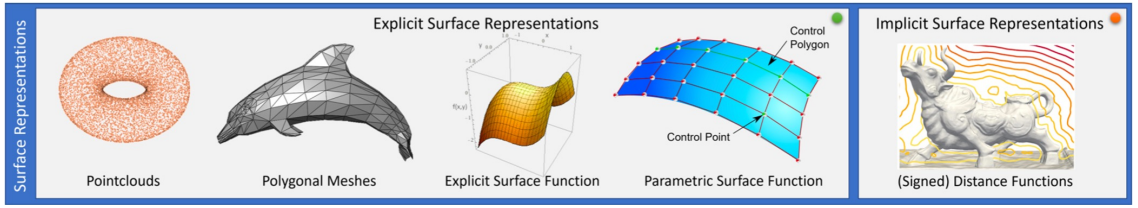
$$S_{implicit} = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \middle| f_{implicit}(x,y,z) = 0 \right\}$$

$$V = \left\{ f_{vol}(x,y,z) \middle| \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \right\}$$



Neural Rendering

- Representing Surfaces



- Why implicit surface representation?
  - Memory efficient
  - Continous
  - High resolution

Neural Rendering

- Representing Volumes

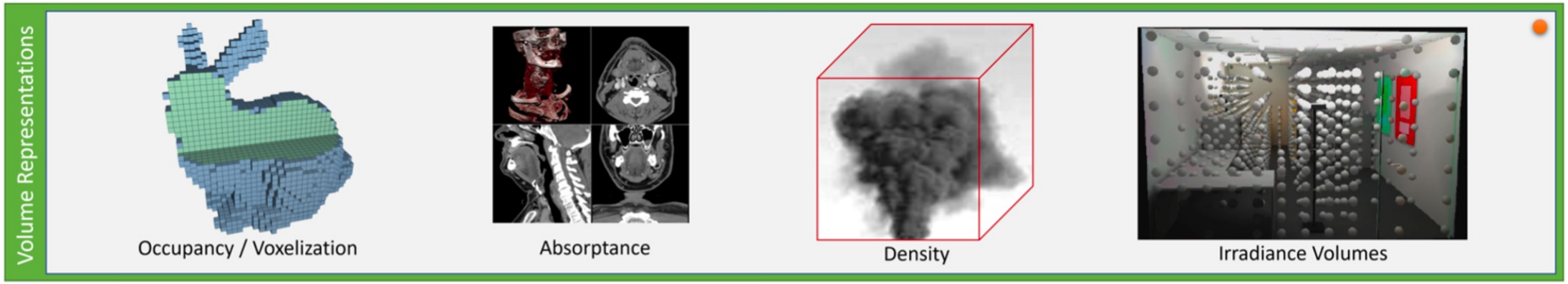
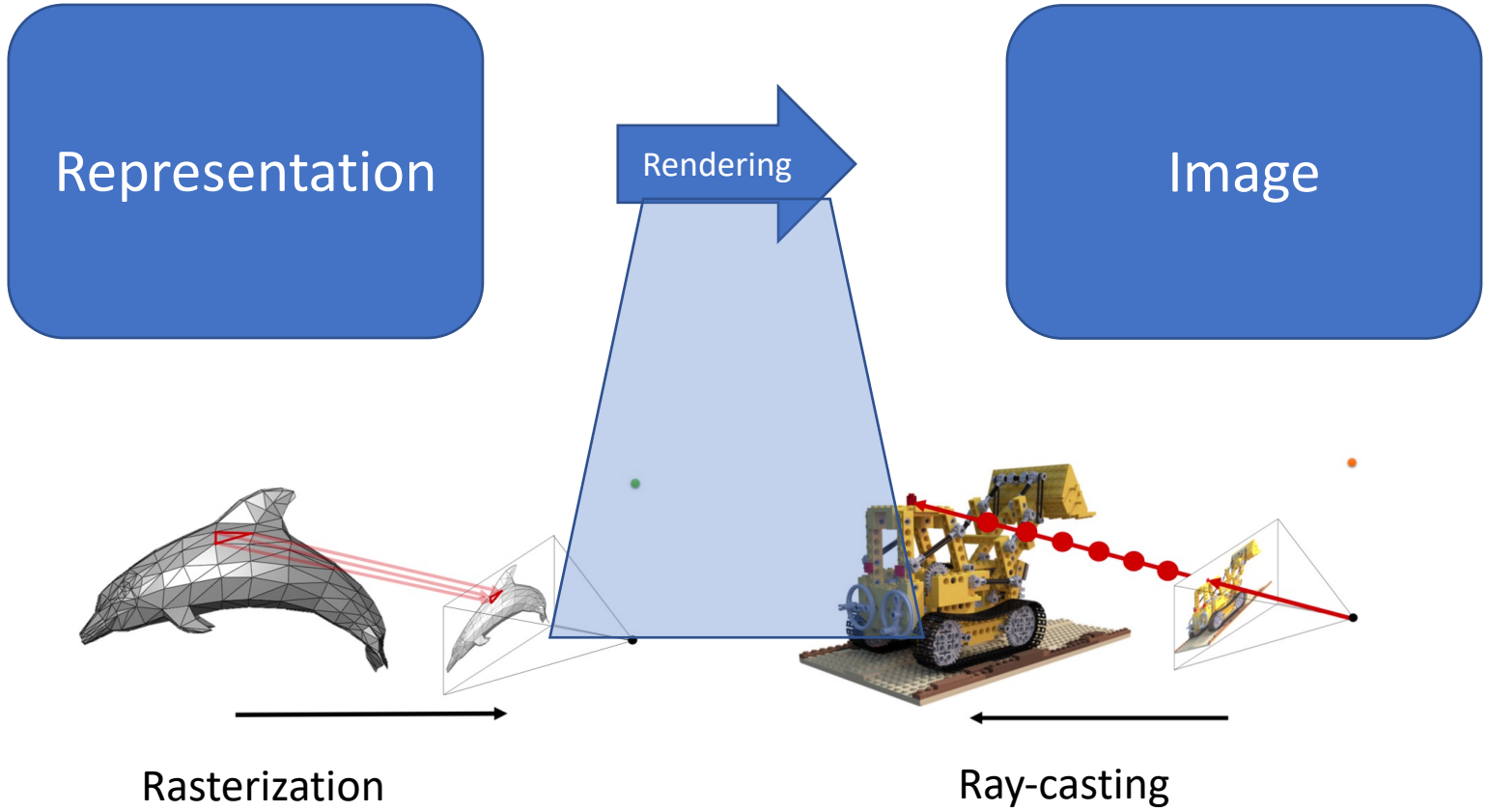


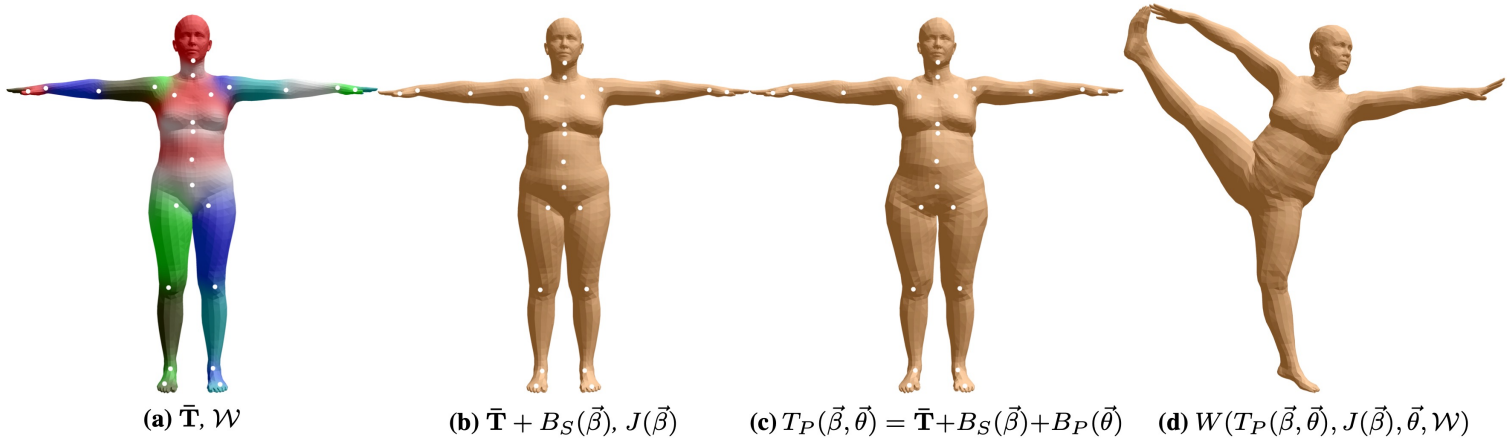
Image Formation

- Rendering



Human Modeling

- SMPL Model



(a) $\bar{\mathbf{T}}, \mathcal{W}$ (b) $\bar{\mathbf{T}} + B_S(\vec{\beta}), J(\vec{\beta})$ (c) $T_P(\vec{\beta}, \vec{\theta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) + B_P(\vec{\theta})$ (d) $W(T_P(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W})$

- Vertices

$$T_P(\vec{\beta}, \vec{\theta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) + B_P(\vec{\theta}) \quad (6)$$

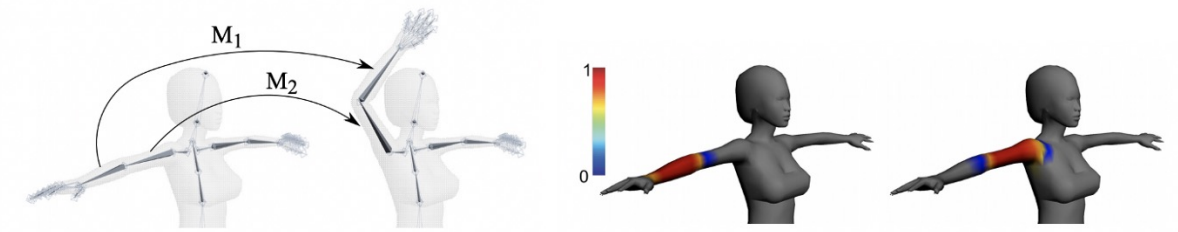
$$\bar{\mathbf{t}}'_i = \sum_{k=1}^{K} w_{k,i} G'_k(\vec{\theta}, J(\vec{\beta}))(\bar{\mathbf{t}}_i + \mathbf{b}_{S,i}(\vec{\beta}) + \mathbf{b}_{P,i}(\vec{\theta})) \quad (7)$$

- Model

$$M(\vec{\beta}, \vec{\theta}) = W(T_P(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W}) \quad (5)$$

Red: Shape
Blue: Pose

- Blend Skinning



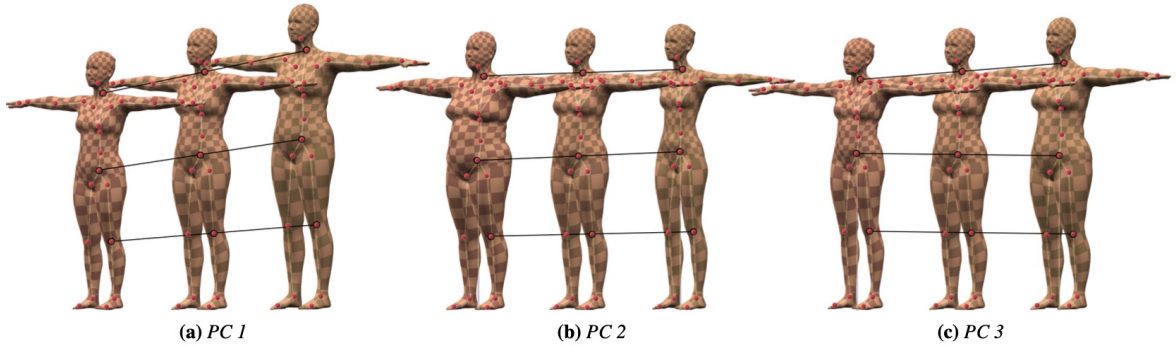
Blend Skinning

- Blend Shapes

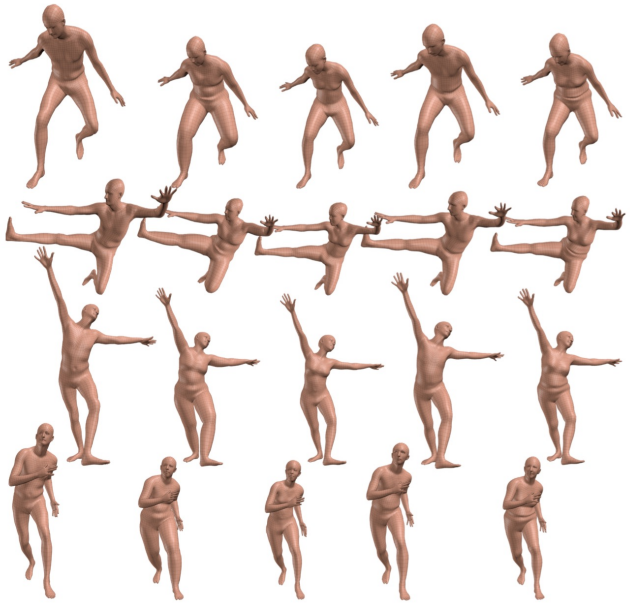
[Pose blend shape]

Human Modeling

- SMPL Parameters
  - Shape (β)



  - Pose (θ)



Human Modeling

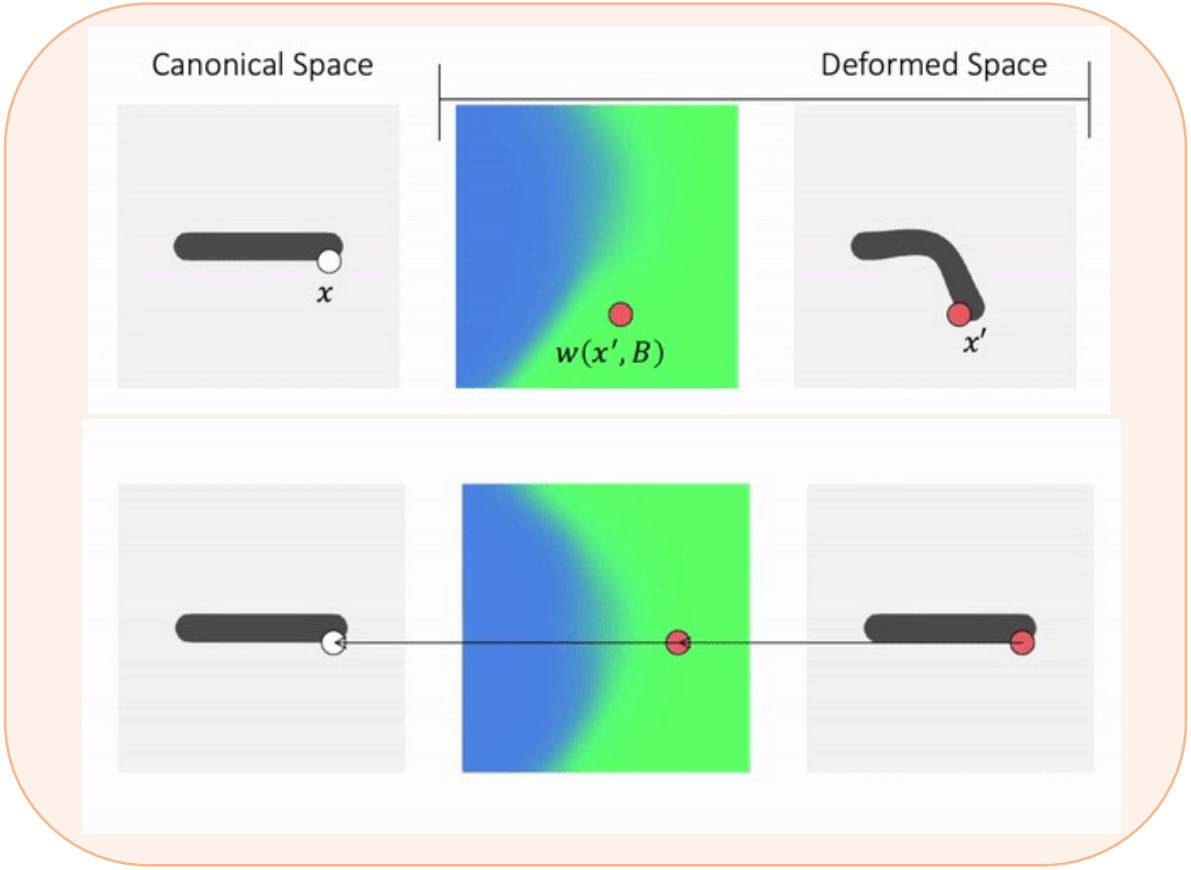
- SNARF : LBS + implicit surface
  - Mesh representation has limitation
    - Resolution-to-memory ratio
    - Fixed topolgy

SNARF results

SMPL model (Mesh)

Human Modeling

- Backward Skinning vs. Forward Skinning

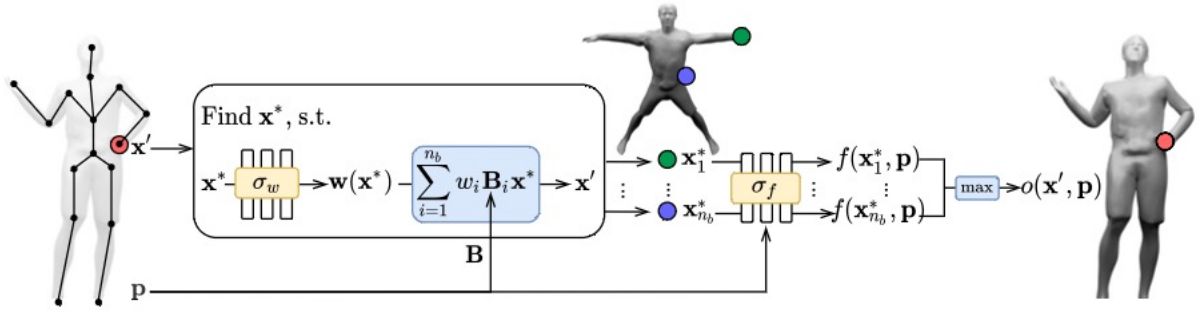


Backward skinning

Forward skinning

Human Modeling

- Modeling overview



- Shape
  - Occupancy probability

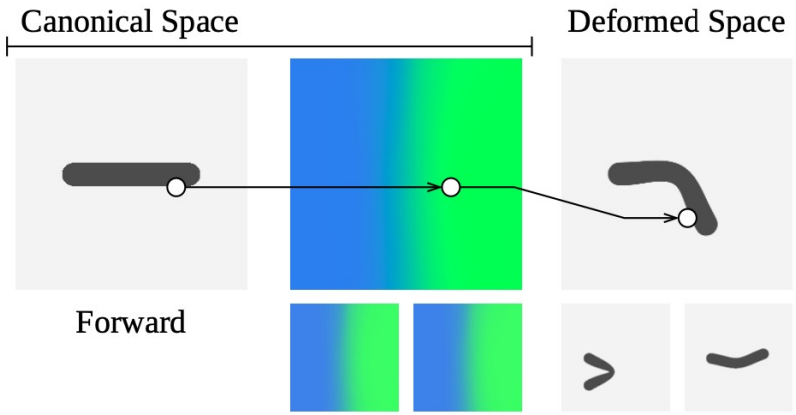
$$f_{\sigma_f} : \mathbb{R}^3 \times \mathbb{R}^{n_p} \rightarrow [0, 1]. \quad (1)$$

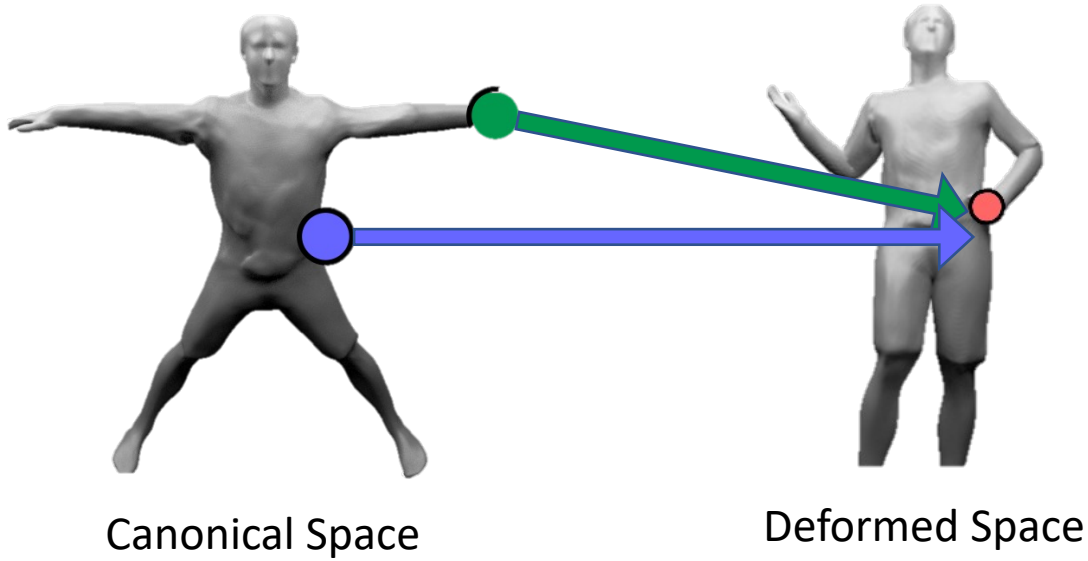
$$\mathcal{S} = \{\mathbf{x} \mid f_{\sigma_f}(\mathbf{x}, \mathbf{p}) = 0.5\}. \quad (2)$$

Result of occupancy network

Human Modeling

- Forward Skinning
  - o(x',p) = f(x*,p)
  - x* 찾기





Human Modeling

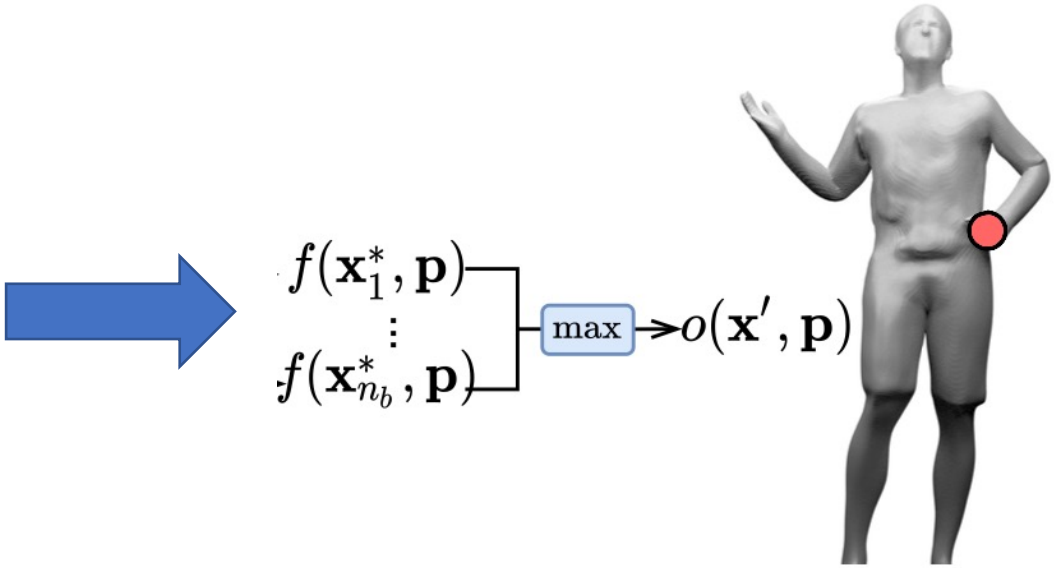
- Correspondence Search

$$\mathbf{d}_{\sigma_w}(\mathbf{x}, \boldsymbol{B}) - \mathbf{x}' = \mathbf{0}, \tag{5}$$

  - No closed-form solution
    - Newton or quasi-Newton methods
  - Handling Multiple Correspondence

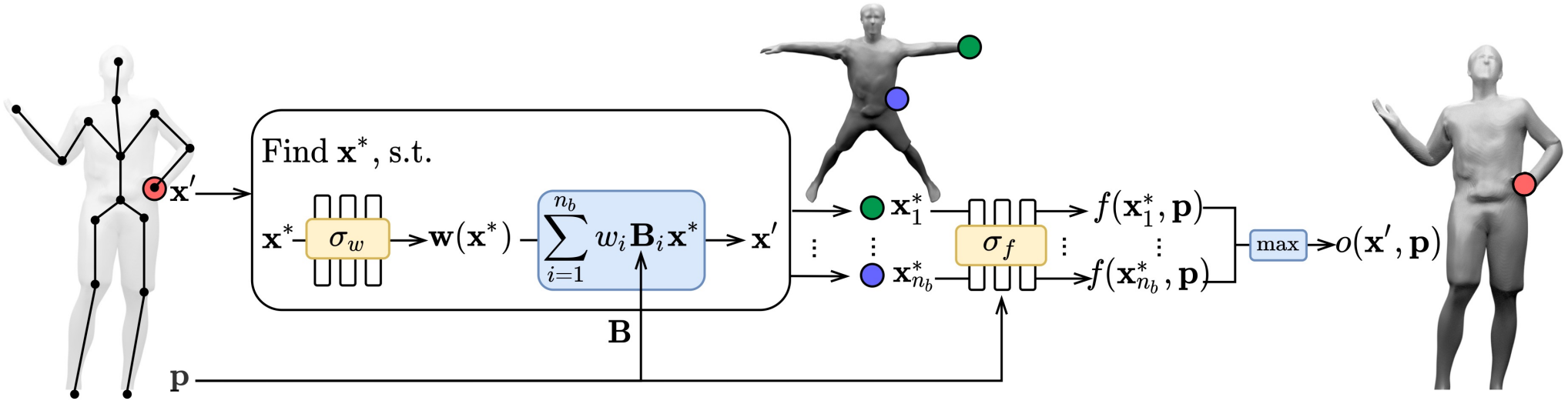
$$\mathcal{X}^* = \{\mathbf{x}_i^* \mid \|\mathbf{d}_{\sigma_w}(\mathbf{x}_i^*, \boldsymbol{B}) - \mathbf{x}'\|_2 < \epsilon\}, \tag{8}$$

$$o(\mathbf{x}', \mathbf{p}) = \max_{\mathbf{x}^* \in \mathcal{X}^*} \{f_{\sigma_f}(\mathbf{x}^*, \mathbf{p})\}. \tag{9}$$



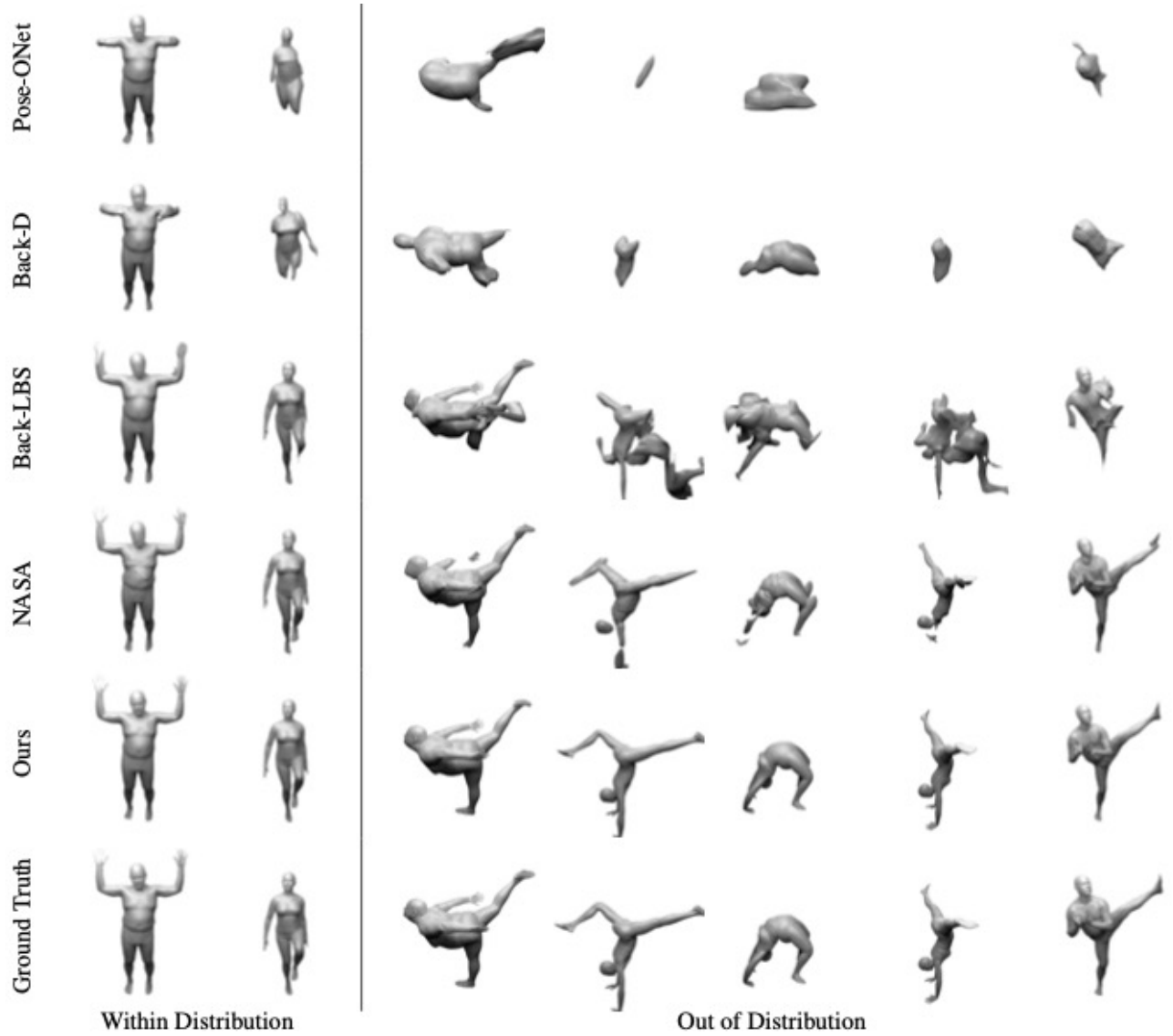
Human Modeling

- Training
  - $\mathcal{L}_{BCE}(o(x', p), o_{gt}(x'))$
- Overview

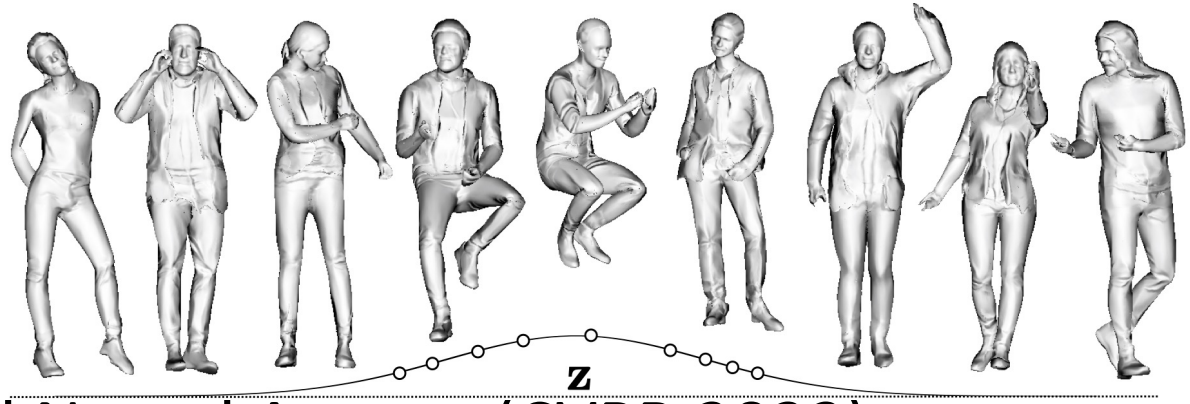


Experiments

- Qualitative Results

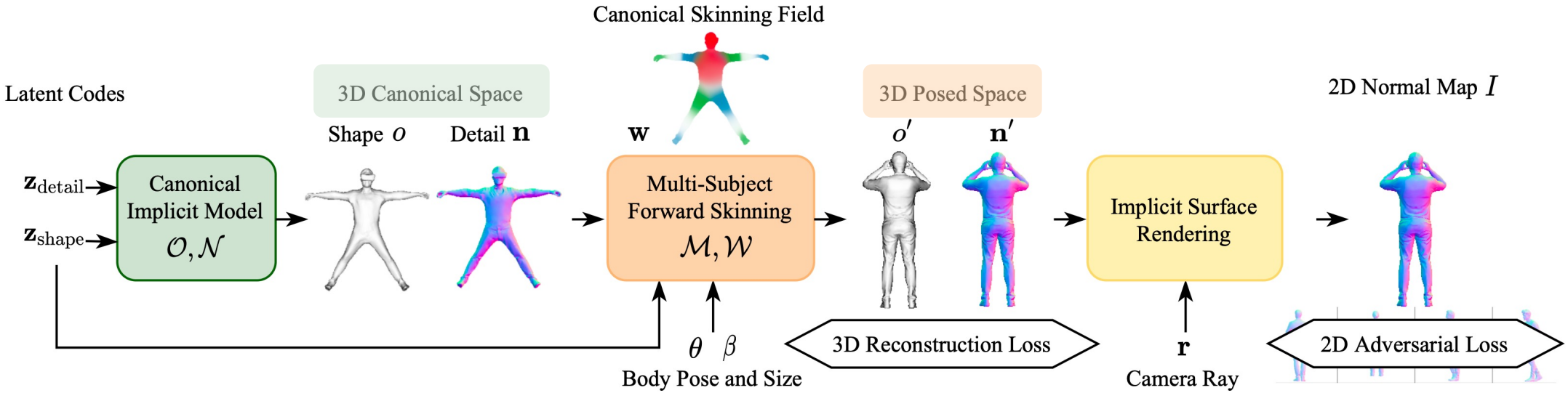


Human Generation

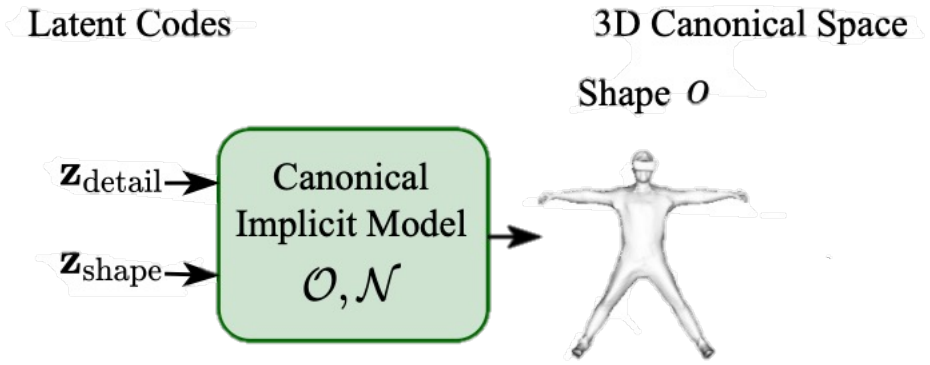


- gDNA: Towards Generative Detailed Neural Avatars (CVPR 2022)

- Overview of method



Human Generation



- Coarse Shape

$$\mathcal{S}(\mathbf{z}_{\text{shape}}) = \{\mathbf{x} \mid \mathcal{O}(\mathbf{x}, \mathbf{z}_{\text{shape}}) = \tau\}, \qquad (1)$$

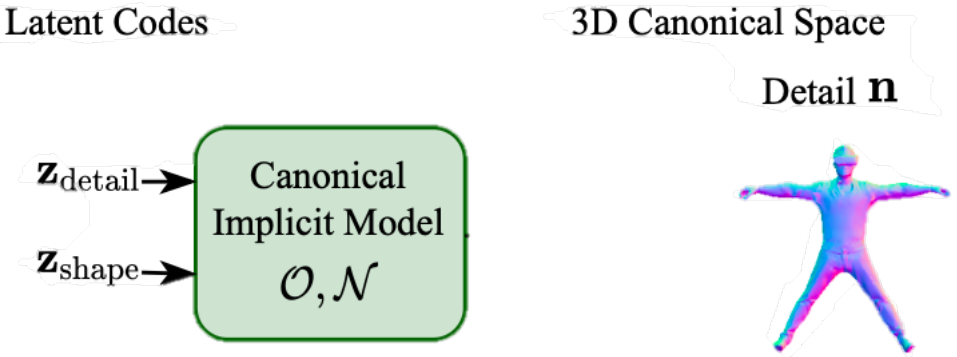
$$\mathcal{O} : \mathbb{R}^3 \times \mathbb{R}^{L_{\text{shape}}} \rightarrow [0,1] \times \mathbb{R}^{L_{\mathbf{f}}} \qquad (2)$$

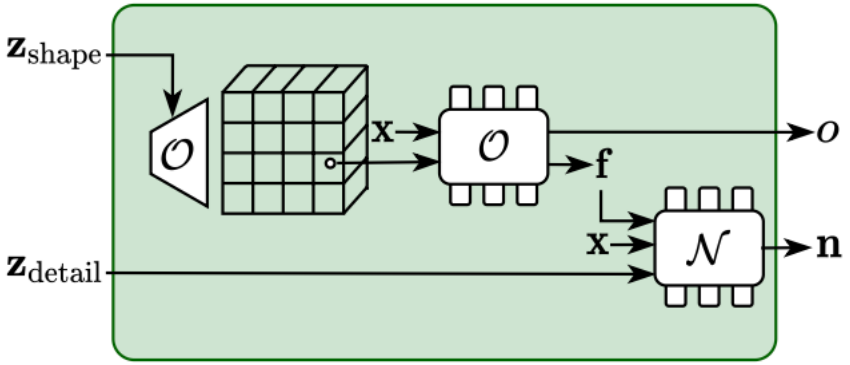
$$(\mathbf{x}, \mathbf{z}_{\text{shape}}) \mapsto (o, \mathbf{f})$$

Human Generation

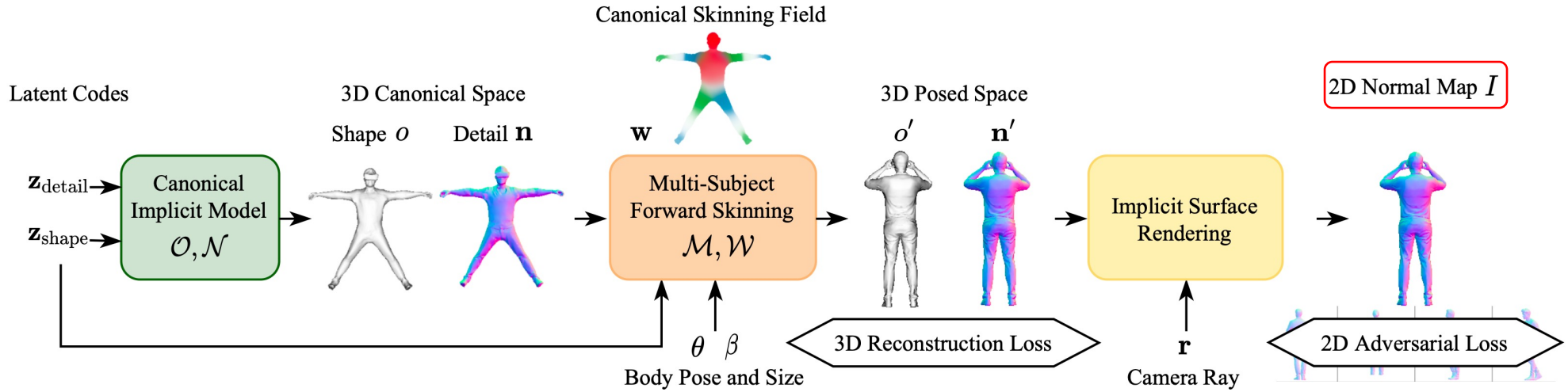
- Detailed Surface Normals

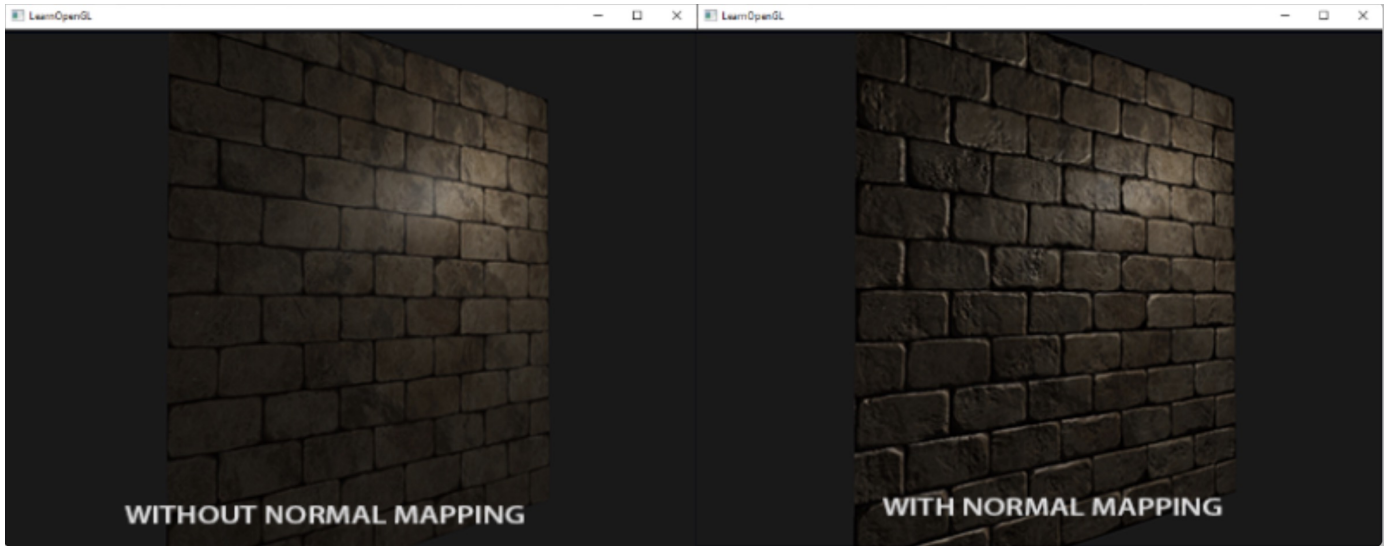
$$\mathcal{N} : \mathbb{R}^3 \times \mathbb{R}^{L_{\text{detail}}} \times \mathbb{R}^{L_{\mathbf{f}}} \rightarrow \mathbb{R}^3 \qquad (3)$$
$$(\mathbf{x}, \mathbf{z}_{\text{detail}}, \mathbf{f}) \mapsto \mathbf{n}$$





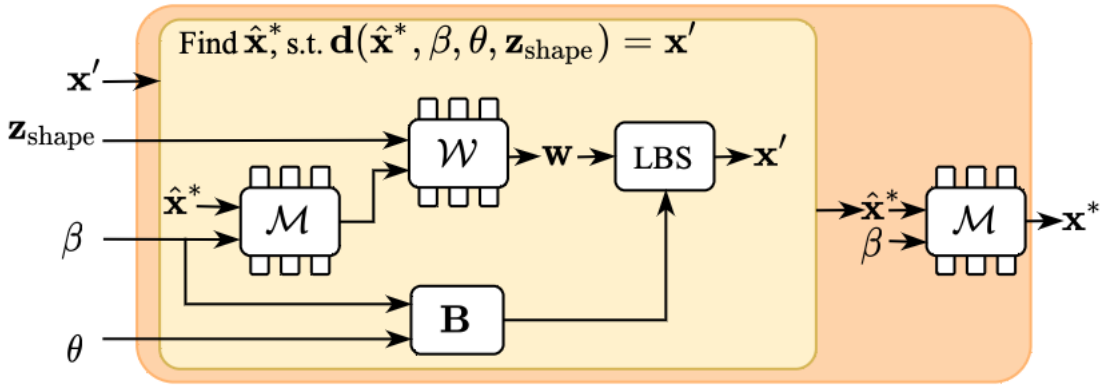
Normal Map


Canonical Skinning Field
Latent Codes
3D Canonical Space
Shape o
Detail $\mathbf{n}$
$\mathbf{w}$
3D Posed Space
o'
$\mathbf{n}'$
2D Normal Map I
$\mathbf{z}_{\text{detail}}$
$\mathbf{z}_{\text{shape}}$
Canonical Implicit Model $\mathcal{O}, \mathcal{N}$
Multi-Subject Forward Skinning $\mathcal{M}, \mathcal{W}$
Implicit Surface Rendering
θ β
Body Pose and Size
3D Reconstruction Loss
$\mathbf{r}$
Camera Ray
2D Adversarial Loss


WITHOUT NORMAL MAPPING
WITH NORMAL MAPPING

Human Generation

- Multi-subject forward skinning

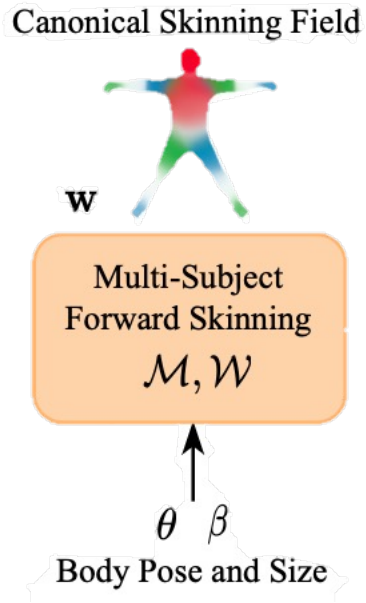


$$\mathcal{W} : \mathbb{R}^3 \to \mathbb{R}^{n_b} \qquad (4)$$
$$\mathbf{x} \mapsto \mathbf{w},$$

Single subject

$$\mathcal{W} : \mathbb{R}^3 \times \mathbb{R}^{L_{\text{shape}}} \to \mathbb{R}^{n_b} \qquad (5)$$
$$(\mathbf{x}, \mathbf{z}_{\text{shape}}) \mapsto \mathbf{w}$$

Multi-subject



$$\mathcal{M} : \mathbb{R}^3 \times \mathbb{R}^{L_\beta} \to \mathbb{R}^3 \qquad (6)$$
$$(\hat{\mathbf{x}}, \boldsymbol{\beta}) \mapsto \mathbf{x}$$

$$\hat{\mathcal{S}}(\mathbf{z}_{\text{shape}}, \boldsymbol{\beta}) = \{\hat{\mathbf{x}} \mid \mathcal{O}(\mathcal{M}(\hat{\mathbf{x}}, \boldsymbol{\beta}), \mathbf{z}_{\text{shape}}) = \tau\} \qquad (7)$$

$$\mathbf{x}' = \mathbf{d}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}_o)$$
$$= \sum_{i=1}^{n_b} \mathcal{W}_i(\mathcal{M}(\hat{\mathbf{x}}, \boldsymbol{\beta}), \mathbf{z}_o) \cdot \boldsymbol{B}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) \cdot \hat{\mathbf{x}}, \qquad (8)$$

Human Generation

- Implicit Differentiable Forward Skinning

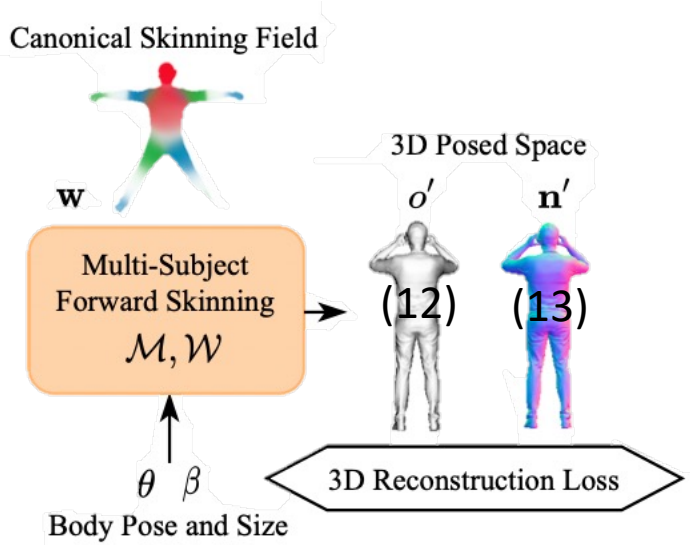
$$\mathbf{d}(\hat{\mathbf{x}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}_{\text{shape}}) - \mathbf{x}' = \mathbf{0}, \tag{9}$$

$$\mathbf{x}^* = \mathcal{M}(\hat{\mathbf{x}}^*, \boldsymbol{\beta}) \tag{10}$$

$$\mathbf{n}' = (\sum_{i=1}^{n_b} \mathcal{W}_i(\mathbf{x}^*, \mathbf{z}_{\text{shape}}) \cdot \mathbf{R}_i)^{-T} \cdot \mathcal{N}(\mathbf{x}^*, \mathbf{f}, \mathbf{z}_{\text{detail}}) \tag{11}$$

$$\mathcal{O}' : (\mathbf{x}', \mathbf{z}_{\text{shape}}, \boldsymbol{\beta}, \boldsymbol{\theta}) \mapsto o', \mathbf{f} \tag{12}$$

$$\mathcal{N}' : (\mathbf{x}', \mathbf{z}_{\text{detail}}, \mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) \mapsto \mathbf{n}' \tag{13}$$

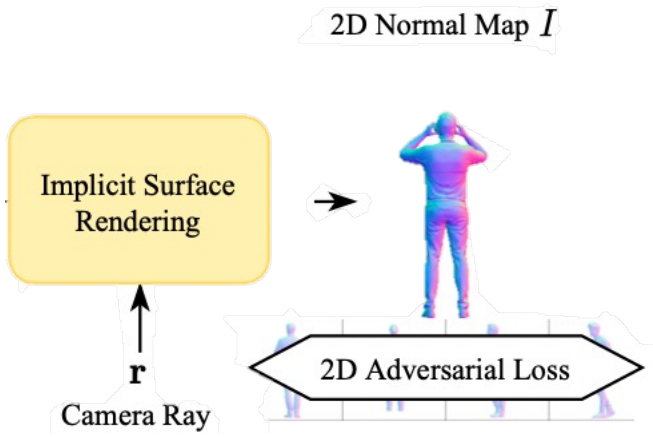


Human Generation

- Implicit Rendering with Skinning
  - Normal map at pixel **p**

$$\mathcal{O}'(\mathbf{x}', \mathbf{z}_{\text{shape}}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \tau, \text{with } \mathbf{x}' = \mathbf{r}_c + t \cdot \mathbf{r}_d \qquad (14)$$

$$I_{\mathbf{p}} = \mathcal{N}'(\mathbf{x}', \mathbf{z}_{\text{detail}}, \mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) \qquad (15)$$



Human Generation

- Training
  - 2-stage training
    - Stage 1: Coarse shape, skinning, warping network
      - $\mathcal{L}_{BCE}\ (\mathcal{O}', o_{gt})$

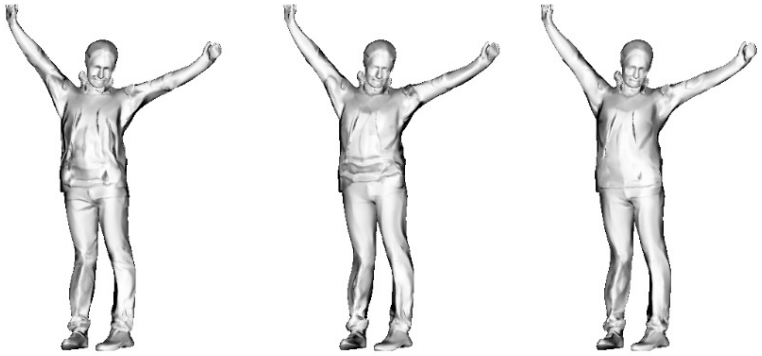
$$\mathcal{L}_{\text{warp}} = \|\mathcal{M}(\mathbf{v}(\boldsymbol{\beta}), \boldsymbol{\beta}) - \mathbf{v}(\boldsymbol{\beta}_0)\|_2^2 \qquad (16)$$

    - Stage 2: Normal network

$$\mathcal{L}_{\text{norm}} = 1 - \mathbf{n}'^{T}_{\text{gt}} \cdot \mathcal{N}'(\mathbf{x}', \mathbf{z}_{\text{detail}}, \mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) \qquad (17)$$

Human Generation

- Experiments
  - Qualitative results



Disentangled Generation
of Shape and Details

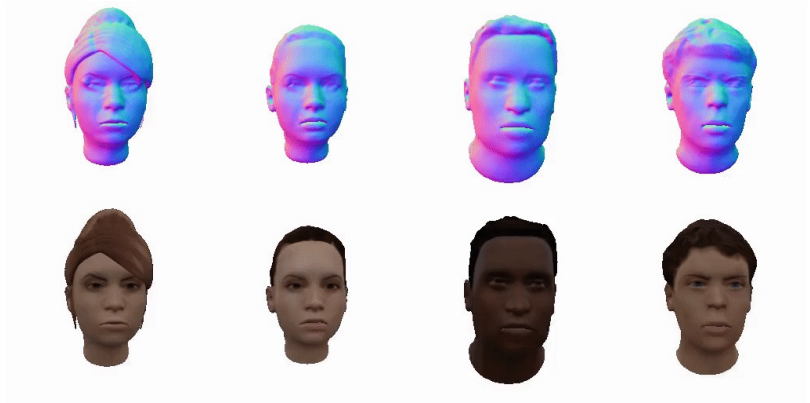
Face Avatar

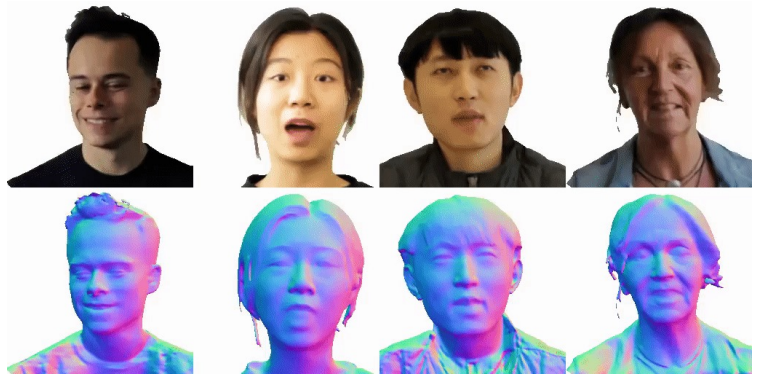
- I M Avatar: Implicit Morphable Head Avatars from Videos

Training video with common expressions

Learned texture, geometry, LBS weight and blendshape fields

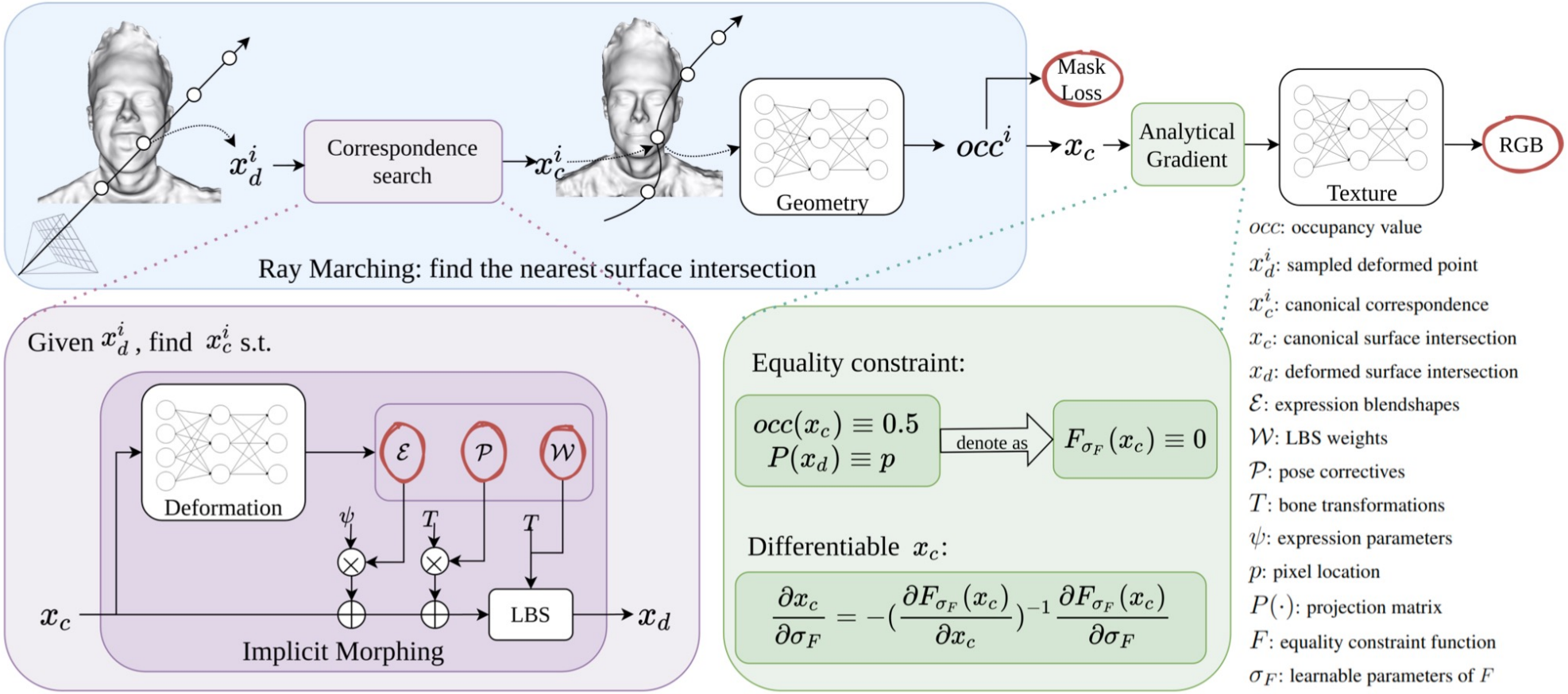
Extrapolation to unseen and extreme expressions





Face Avatar

- Method Overview

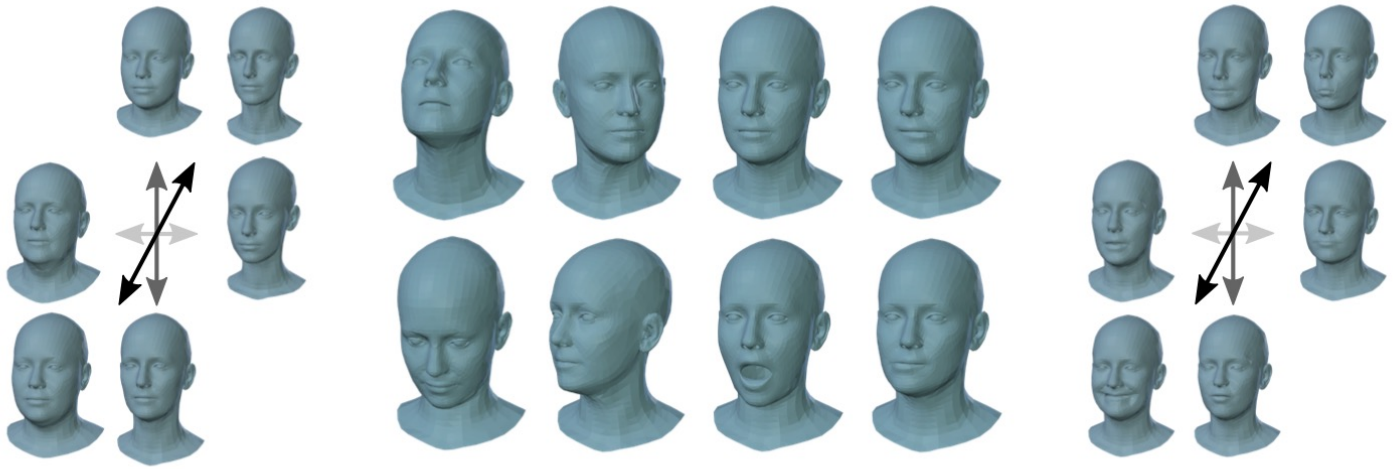


Face Avatar

- FLAME Face Morphable Model
  - Face model analogous to SMPL model (body)

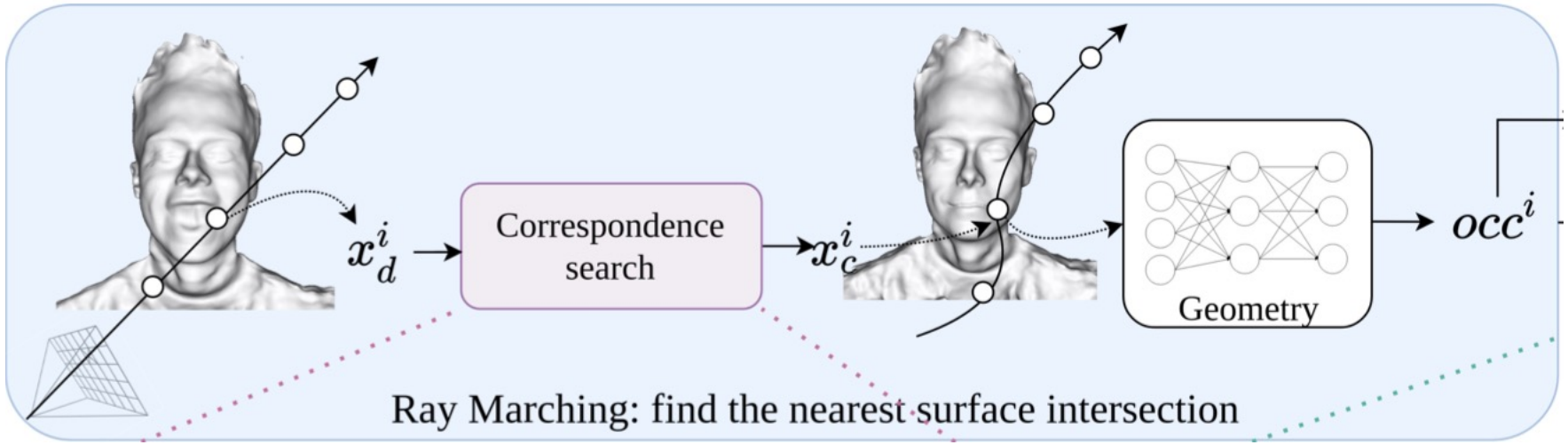
$$M(\boldsymbol{\theta}, \boldsymbol{\psi}) = LBS(T_P(\boldsymbol{\theta}, \boldsymbol{\psi}), J(\boldsymbol{\psi}), \boldsymbol{\theta}, \mathcal{W}), \qquad (1)$$

$$T_P(\boldsymbol{\theta}, \boldsymbol{\psi}) = \bar{\boldsymbol{T}} + B_E(\boldsymbol{\psi}; \mathcal{E}) + B_P(\boldsymbol{\theta}; \mathcal{P}), \qquad (2)$$



Face Avatar

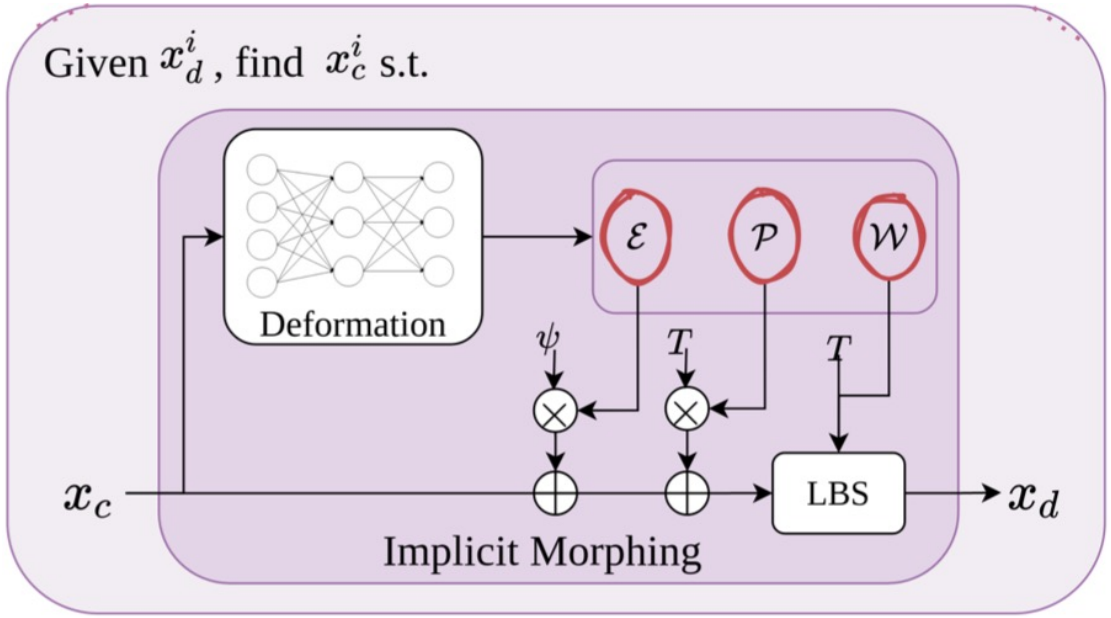
- I M Avatar
  - Geometry



$$f_{\sigma_f}(x, l) : \mathbb{R}^3 \times \mathbb{R}^{n_l} \rightarrow occ. \qquad (3)$$

Face Avatar

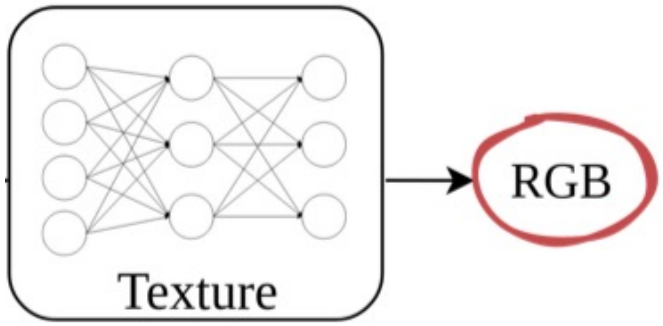
- Deformation



$$d_{\sigma_d}(x) : \mathbb{R}^3 \rightarrow \mathcal{E}, \mathcal{P}, \mathcal{W}. \quad (4)$$

$$x_d = LBS(x_c + B_P(\boldsymbol{\theta}; \mathcal{P}) + B_E(\boldsymbol{\psi}; \mathcal{E}), J(\boldsymbol{\psi}), \boldsymbol{\theta}, \mathcal{W}), \quad (5)$$

Face Avatar

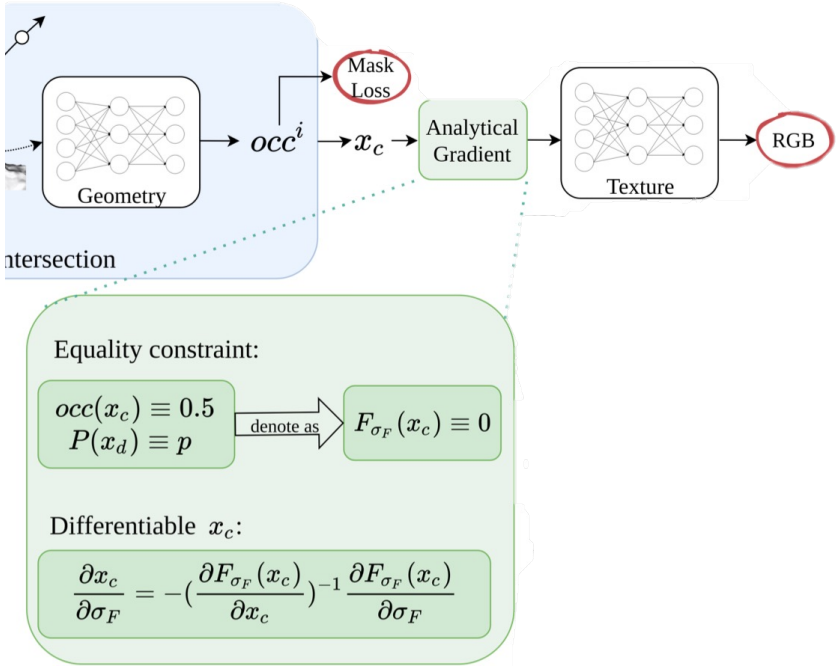


- Normal-conditioned texture
  - Normalized gradient of occupancy field

$$\frac{\partial f_{\sigma_f}(x_c)}{\partial x_d} = \frac{\partial f_{\sigma_f}(x_c)}{\partial x_c}\frac{\partial x_c}{\partial x_d} = \frac{\partial f_{\sigma_f}(x_c)}{\partial x_c}\left(\frac{\partial w_{\sigma_d}(x_c)}{\partial x_c}\right)^{-1}. \tag{6}$$

Face Avatar

- Differentiable rendering
  - Non-rigid ray marching
  - Gradient



$$\frac{dF_{\sigma_F}(x_c)}{d\sigma_F} = 0$$

$$\Leftrightarrow \quad \frac{\partial F_{\sigma_F}(x_c)}{\partial \sigma_F} + \frac{\partial F_{\sigma_F}(x_c)}{\partial x_c}\frac{\partial x_c}{\partial \sigma_F} = 0 \qquad (10)$$

$$\Leftrightarrow \quad \frac{\partial x_c}{\partial \sigma_F} = -(\frac{\partial F_{\sigma_F}(x_c)}{\partial x_c})^{-1}\frac{\partial F_{\sigma_F}(x_c)}{\partial \sigma_F}.$$

$$f_{\sigma_f}(x_c) \equiv 0.5, \qquad (8)$$

$$(w_{\sigma_d}(x_c) - r_o) \times r_d \equiv 0, \qquad (9)$$

Face Avatar

- Training Objectives
  - RGB Loss (Eq. (11))
    - Pixel color
  - Mask Loss (Eq. (12))
  - FLAME Loss (optional, Eq. (13))
    - Use prior knowledge from FLAME

$$\mathcal{L}_{RGB} = \frac{1}{|P|} \sum_{p \in P^{in}} \|C_p - c_{\sigma_c}(x_c)\|_1, \tag{11}$$

$$\mathcal{L}_{M} = \frac{1}{|P|} \sum_{p \in P \setminus P^{in}} CE(O_p, f_{\sigma_f}(x_c^\star)), \tag{12}$$

$$\mathcal{L}_{FL} = \frac{1}{|P|} \sum_{p \in P^{in}} (\lambda_e \|\mathcal{E}_p^{GT} - \mathcal{E}_p\|_2 + \lambda_p \|\mathcal{P}_p^{GT} - \mathcal{P}_p\|_2 + \lambda_w \|\mathcal{W}_p^{GT} - \mathcal{W}_p\|_2), \tag{13}$$

Face Avatar

- Experiments

Method	Expression ↓	L_1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
C-Net	3.615	0.05824	22.23	0.9524	0.03421
D-Net	3.769	0.06130	21.77	0.9474	0.03227
B-Morph	2.786	0.04980	23.50	0.9599	0.02231
Fwd-Skin	3.088	0.05456	22.92	0.9586	0.02781
NerFACE [17]	2.994	**0.04564**	23.58	0.9596	0.02156
Ours-	2.843	0.04918	23.68	0.9615	0.02155
Ours	**2.548**	0.04878	**23.91**	**0.9655**	**0.02085**

References

[1] Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., ... & Zollhöfer, M. (2020, May). State of the art on neural rendering. In *Computer Graphics Forum* (Vol. 39, No. 2, pp. 701-727).

[2] Chen, X., Zheng, Y., Black, M. J., Hilliges, O., & Geiger, A. (2021). SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11594-11604).

[3] Chen, X., Jiang, T., Song, J., Yang, J., Black, M. J., Geiger, A., & Hilliges, O. (2022). gDNA: Towards Generative Detailed Neural Avatars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20427-20437).

[4] Zheng, Y., Abrevaya, V. F., Bühler, M. C., Chen, X., Black, M. J., & Hilliges, O. (2022). Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13545-13555).