NeuralSMPL: 3D-aware Neural Body Fitting for Robust Human Pose Estimation under Occlusion

Yi Zhang, Adam Kortylewski, Jieru Mei, Angtian Wang, Alan Yuille

Robustness of Objects Model to Occlusion

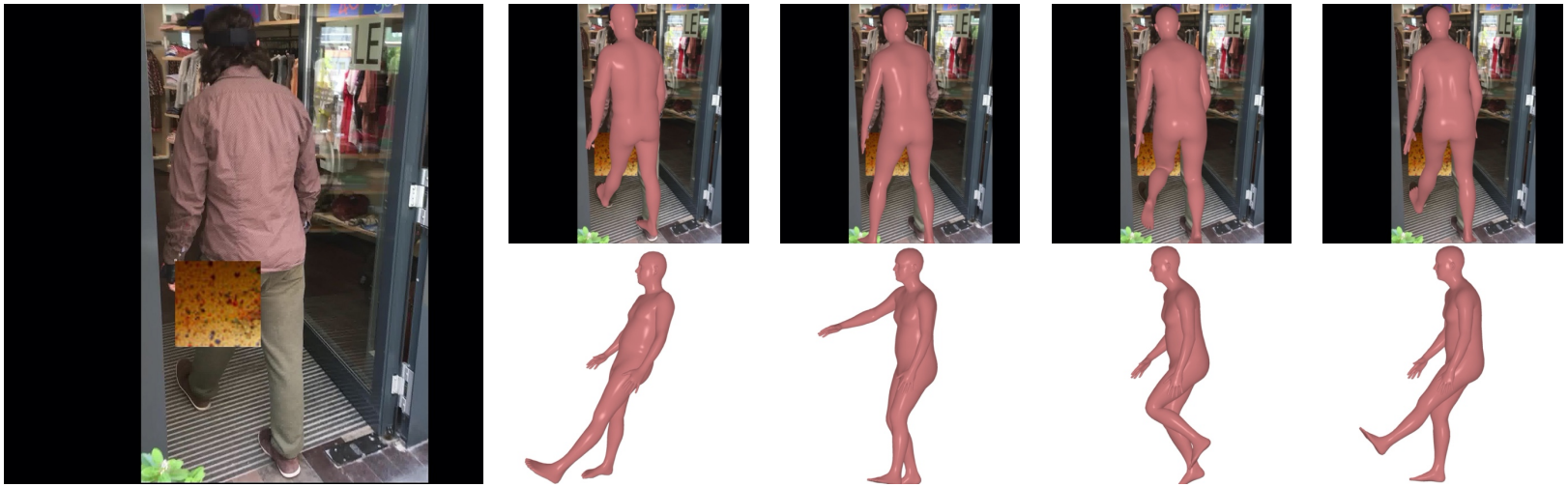
- A goal of BRIAR is to develop models of humans which are robust to challenging nuisance factors.
- We focus on robustness to occlusion.
- Our previous work on rigid object categories shows that we can develop generative models of objects which are robust to occlusion. This was done for object classification (A. Kortylewski et al.) and for 3D pose estimation (NeMo).
- We present work which extends this approach to estimating the 3D configuration of humans under occlusion.

Main Points

- Standard regression methods (e.g., Deep Networks) do not generalize well to occlusions. They overgeneralize to context.
- We represent humans by mesh models of vertices SMPL. We learn a generative model of feature vectors which we call NeuralSMPL. This performs as well as Deep Nets if there is no occlusion but outperforms them if there is occlusion.
- NeuralSMPL specifies a generative model of feature vectors. This model is conditioned on the orientation of the limbs and the viewpoint of the object. The model is factorized over different vertices of the mesh model. An outlier process is added to enable robustness to occlusion (without any training on occluded data).
- Inference is performed by gradient descent. The energy landscape is smooth (since the model is generative on feature vectors).

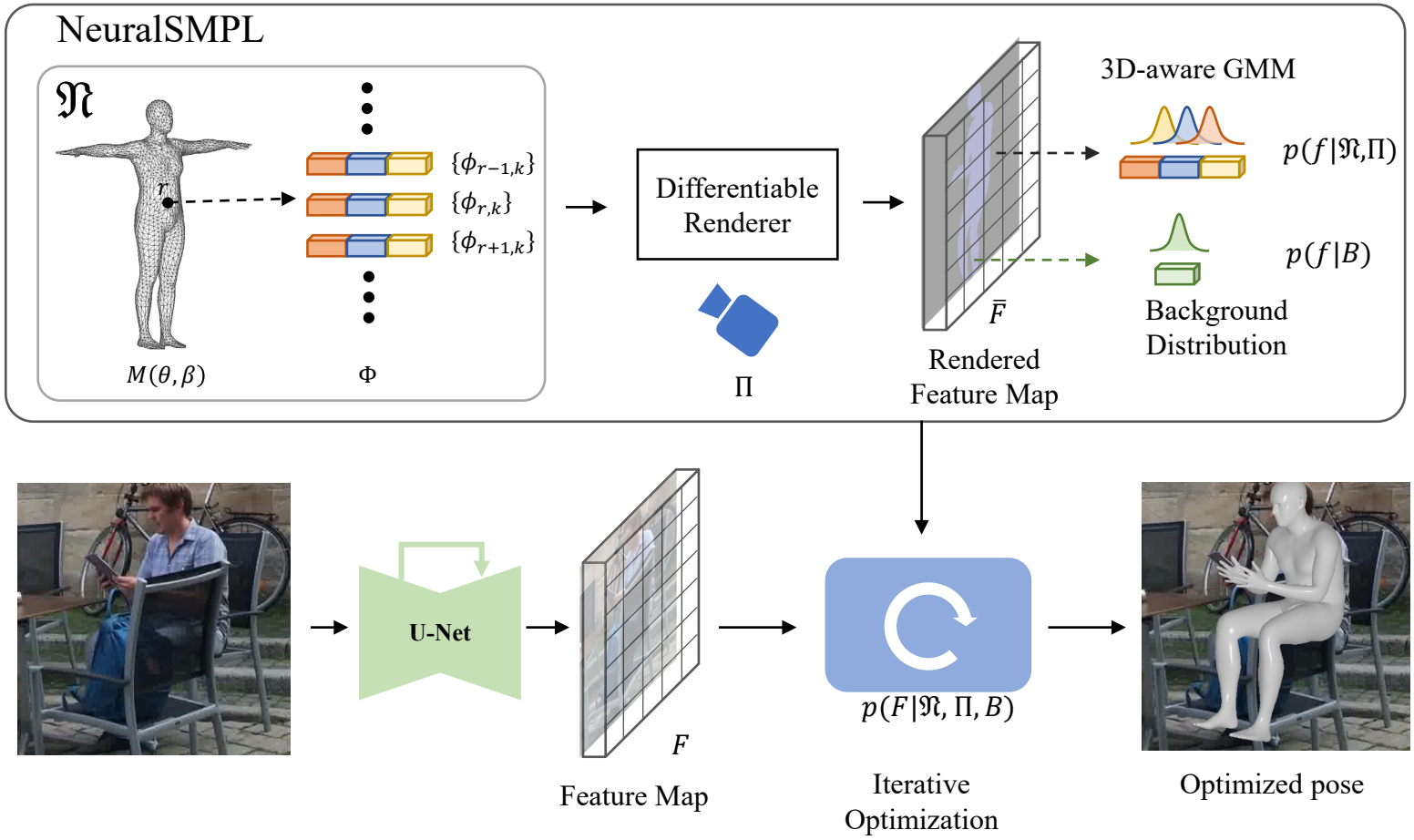
Motivation: Human Pose Estimation under Occlusion

- Deep Net methods have limited robustness to occlusion, because they are global and overexploit context.
- NeuralSMPL is local and has an occlusion process, making it robust to occlusion and competitive to Deep Nets if there is no occlusion.

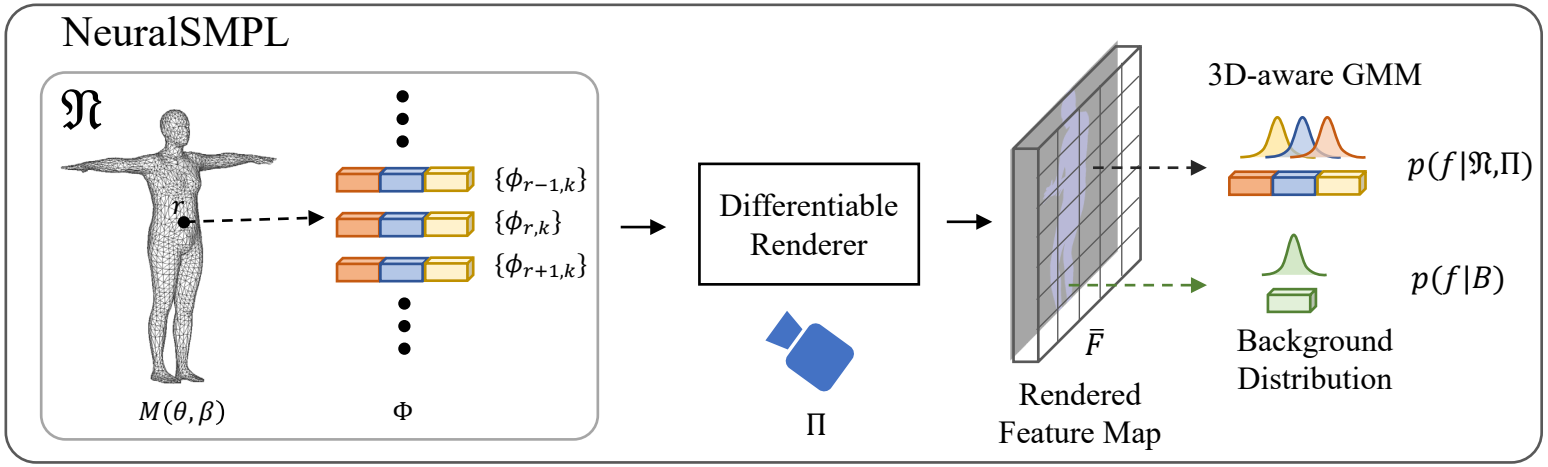


(a) Input Image (b) GT (c) Regression (d) Regression+ Model-Fitting (e)

3D Human Pose Estimation by NeuralSMPL



NeuralSMPL: A 3D-aware Generative Model of Neural Textures



- NeuralSMPL consists of a neural mesh model $\mathfrak{N}(M(\theta,\beta),\Phi)$ and a distribution of feature vectors in image space. Local and Factorizable.

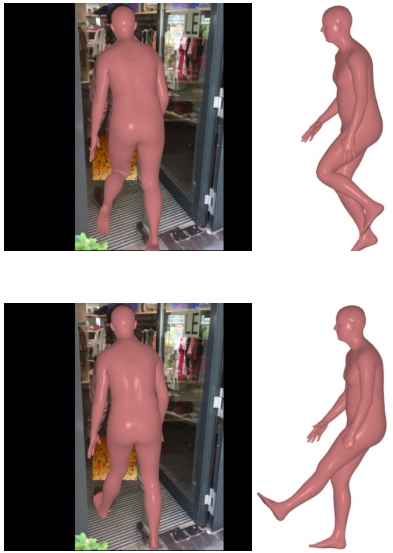
$$p(F|\mathfrak{N},\Pi,B)=\prod_{i\in\mathcal{FG}}p(f_i|\mathfrak{N})\prod_{i'\in\mathcal{BG}}p(f_{i'}|B), \qquad (1)$$

- NeuralSMPL is naturally robust to occlusion, without training on occluded data,
- by Adding an Outlier Process:

$$p(F|\mathfrak{N},\Pi,B,z_i)=\prod_{i\in\mathcal{FG}}\left[p(f_i|\mathfrak{N})p(z_i{=}1)\right]^{z_i}\left[p(f_i|B)p(z_i{=}0)\right]^{(1-z_i)}\prod_{i'\in\mathcal{BG}}p(f_{i'}|B), \qquad (2)$$

3D-aware GMMs

- We use Gaussian Mixture Models (GMMs) to explicitly encode 3D information into features ϕ
- Each mixture focuses on modeling a subspace of 3D orientation of the limb that the vertex belongs to

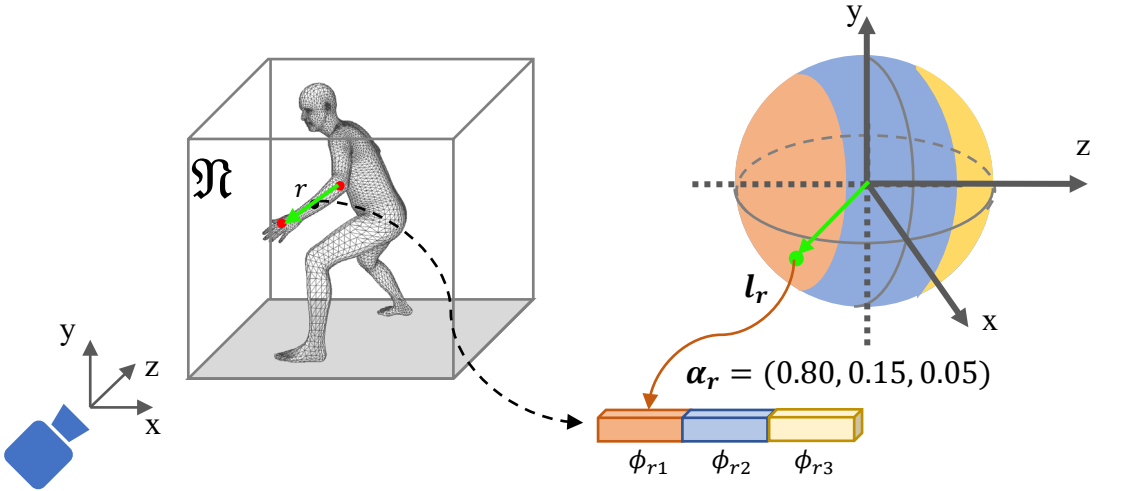


$$p(f_i|\mathfrak{N}, \mathit{\Pi}) = \sum_{k=1}^{K} \alpha_{rk} \mathcal{N}(f_i|\phi_{rk}, \sigma_{rk}^2 I)$$

$$p(f_i|\mathfrak{N}(\Phi), \Pi) = \sum_{k=1}^{K} \alpha_{rk} \mathcal{N}(f_i|\phi_{rk}, \sigma_{rk}^2 I)$$

The correspondence between vertex ‘r’ and pixel ‘i’ is determined by the NeuralSMPL model $\mathfrak{N}$ and the camera parameter Π

3D-Aware Gaussian Mixtures



$$p(f_r|\mathfrak{N},\Pi) = 0.80 * \mathcal{N}(\phi_{r1}, \sigma^2 I) + 0.15 * \mathcal{N}(\phi_{r2}, \sigma^2 I) + 0.05 * \mathcal{N}(\phi_{r3}, \sigma^2 I)$$

$$\alpha_{rk} = sigmoid(\tau(\gamma_r - \frac{k\pi}{K})) + sigmoid(-\tau(\gamma_r - \frac{(k+1)\pi}{K})) - 1 \qquad (6)$$

$$\gamma_r = \arccos(\mathbf{l}_r \cdot \mathbf{z}_+)$$

3D-Aware Gaussian Mixtures

- Soft assignment for differentiability

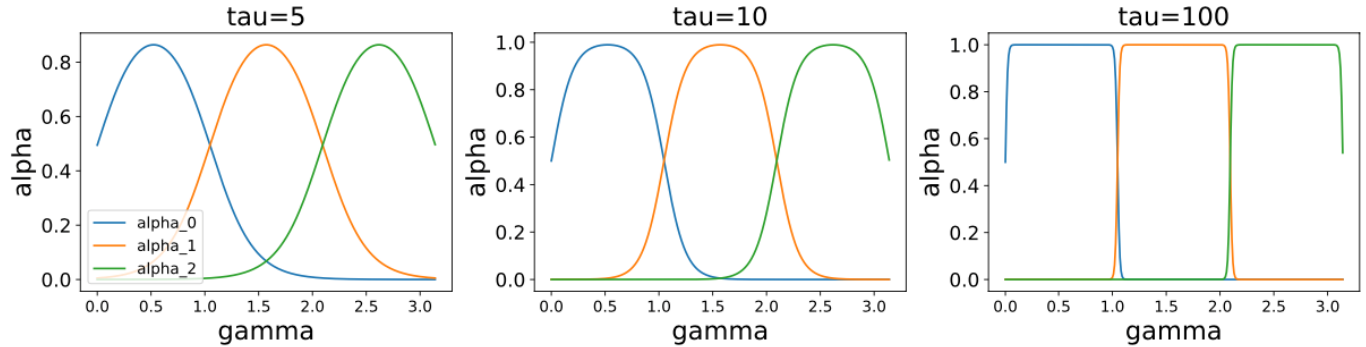


Fig. 2: Approximate the rectangular assignment function with sigmoid functions. τ controls the softness of the approximation.

Training of NeuralSMPL

- Image feature extractor
  - 3D-aware contrastive learning
- $\mathcal{L}\_Vertex$ encourages features of different vertices to be distinct from each other
- $\mathcal{L}\_3D$ encourages features of the same vertex in different 3D poses to be different

$$\mathcal{L}_{Vertex}(F, \mathcal{FG}) = -\sum_{i \in \mathcal{FG}} \sum_{i' \in \mathcal{FG} \setminus \{i\}} \|f_i - f_{i'}\|^2$$

$$\mathcal{L}_{3D}(F, \mathcal{FG}) = -\sum_{r} \sum_{k} \sum_{k' \in K \setminus \{k\}} \|f_{rk} - f_{rk'}\|^2$$

$$\mathcal{L}_{BG}(F, \mathcal{FG}, \mathcal{BG}) = -\sum_{i \in \mathcal{FG}} \sum_{j \in \mathcal{BG}} \|f_i - f_j\|^2$$

- NeuralSMPL parameters Φ
  - Maximum likelihood estimation (MLE)
- Minimizing the negative log-likelihood of feature representations over the whole training set.
- $\{f_j\}$ are all feature vectors that correspond to vertex r, which is obtained from the ground truth pose and camera parameters.

$$\phi_{rk} = \frac{\sum_j \alpha_{jrk} f_j}{\sum_j \alpha_{jrk}},$$

Inference with Multitask Integration

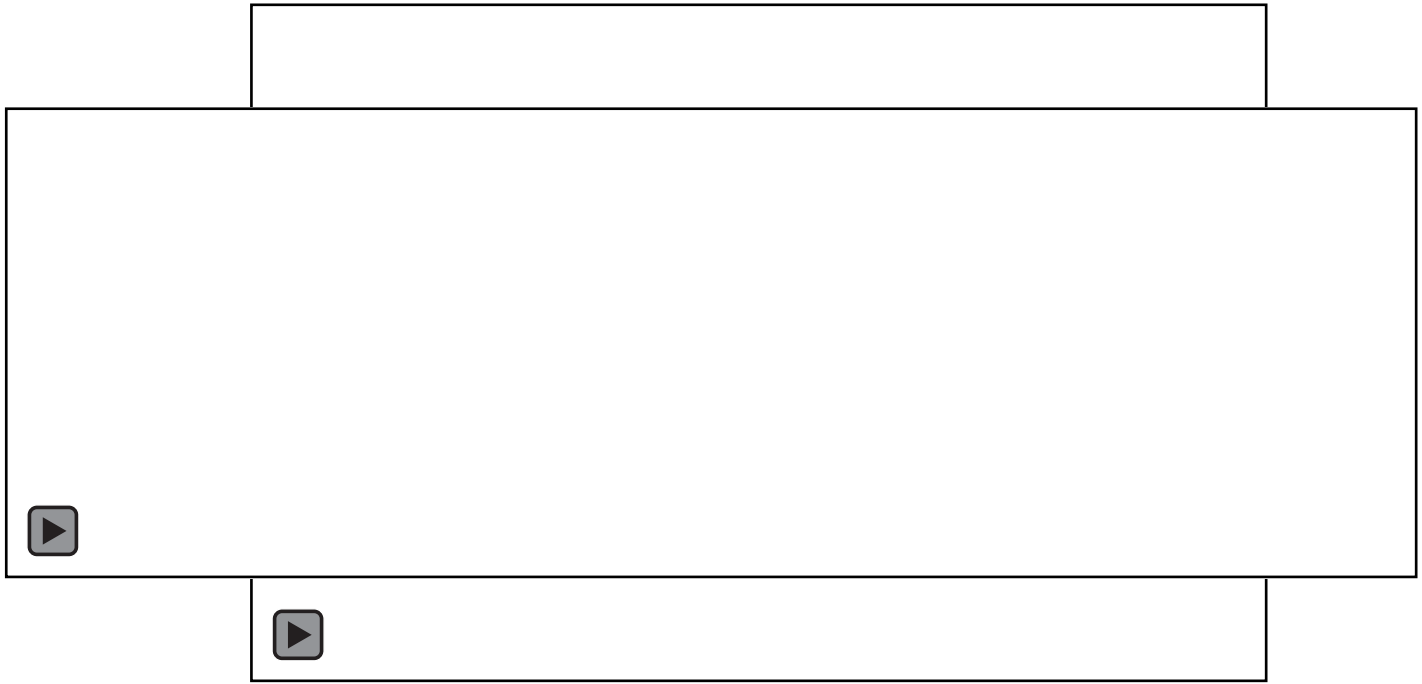
- Incorporate auxiliary losses for better convergence:
  - 2D keypoints reprojection loss
  - Part segmentation loss
  - 3D pose prior – a VAE prior (VPoser)
  - Algorithm is Gradient Descent using Adam optimizer. A Deep Net or any other pose estimation method can be used to initialize.

$$\mathcal{L}_{inference}(F_j, \mathfrak{N}, \Pi, B) = \mathcal{L}_{NLL}(F_j, \mathfrak{N}, \Pi, B) + \mathcal{L}_{reproj}(\hat{J}_{2D}, \mathfrak{N}, \Pi) + \mathcal{L}_{partseg}(\hat{P}, \mathfrak{N}, \Pi) + \mathcal{L}_{prior}(\mathfrak{N}).$$

Experiments

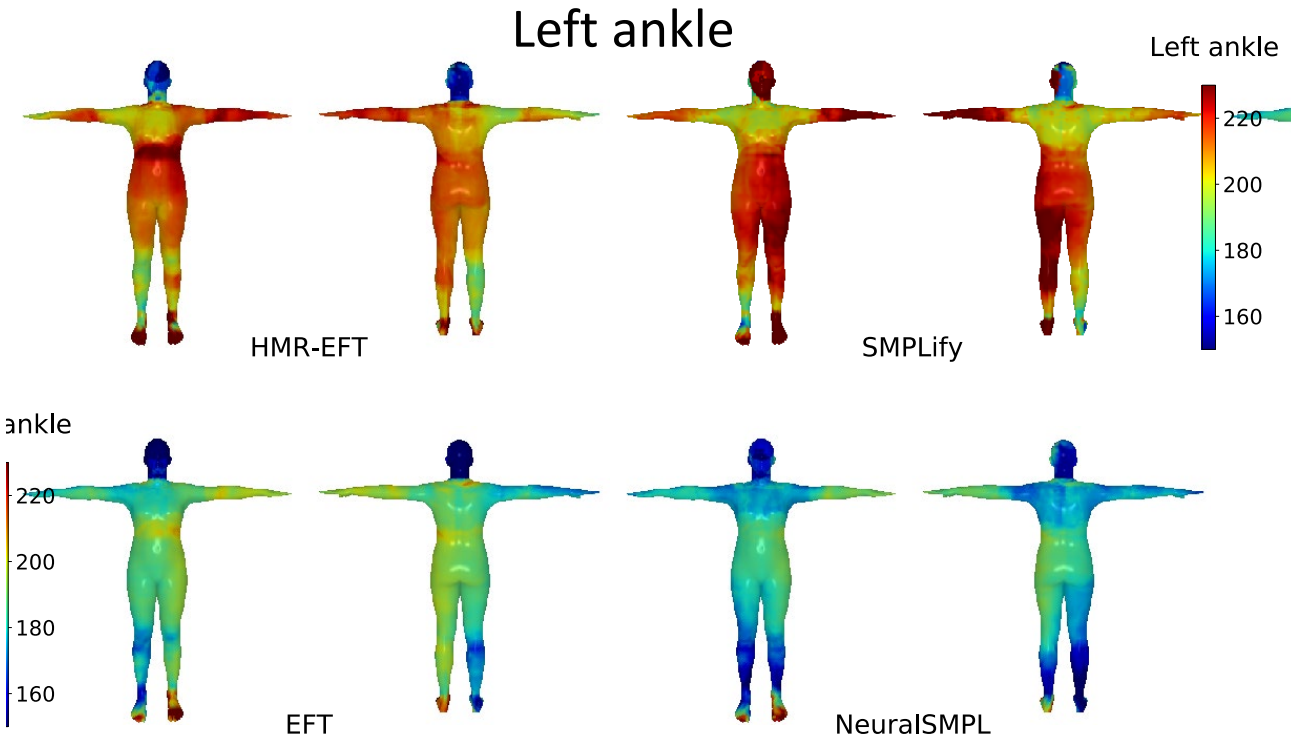
Adversarial Occlusion Robustness Evaluation

- 3DPW-AdvOcc protocol
  - Slide an occlusion patch to find the worst prediction.



Occlusion Sensitivity Mesh

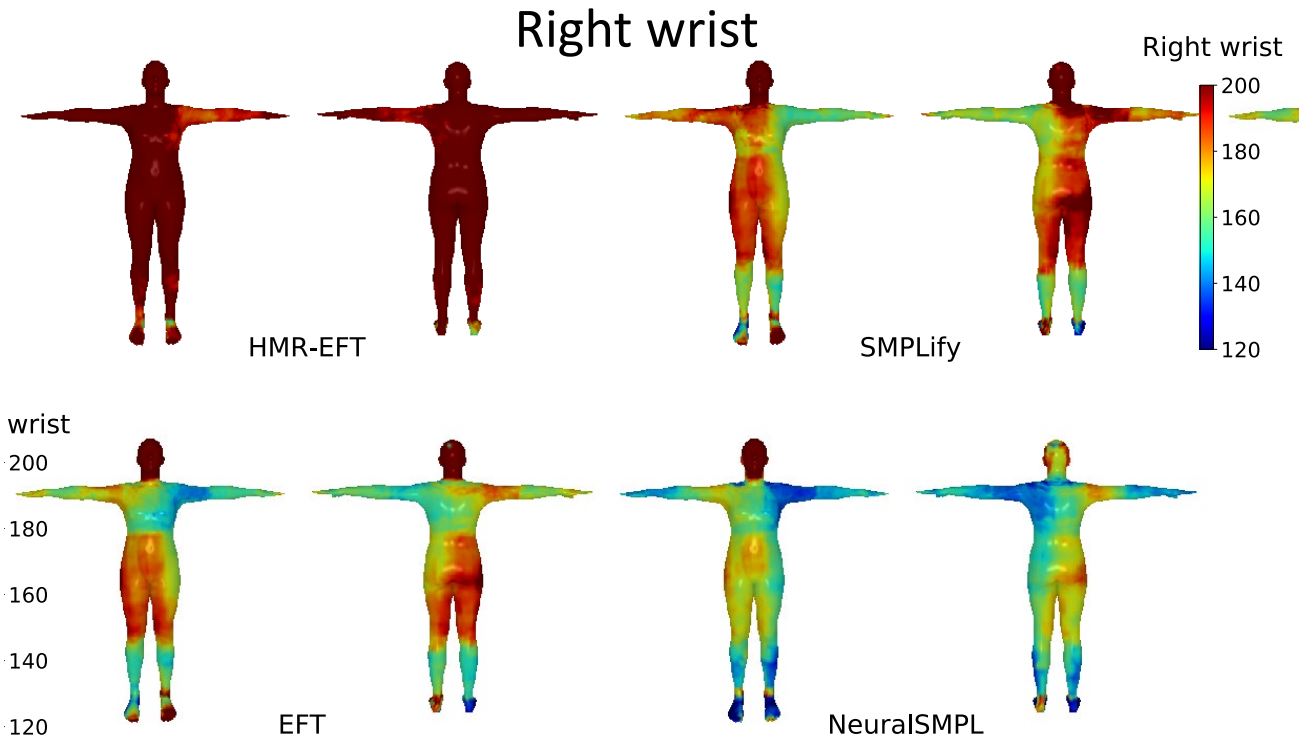
- Vertex color encodes the average MPJPE of the joint when the vertex is occluded



HMR-EFT is the regressor. SMPLify and EFT are optimization-based methods.

Occlusion Sensitivity Mesh

- Vertex color encodes the average MPJPE of the joint when the vertex is occluded



HMR-EFT is the regressor. SMPLify and EFT are optimization-based methods.

Compare to SotA Regression-based Methods

- Methods are trained on the same datasets and use the same ResNet50 backbone

Table 1: **Performance on 3DPW and 3DPW-AdvOcc.** NeuralSMPL outperforms state-of-the-art regression-based methods on 3DPW-AdvOcc while being on par or better on 3DPW. Evaluation metrics reported: MPJPE (mm, ↓ the better), PA-MPJPE (mm, ↓ the better), and PCKh (%, ↑ the better).

Method	3DPW			3DPW-AdvOcc@40			3DPW-AdvOcc@80		
	MPJPE	PA-MPJPE	PCKh	MPJPE	PA-MPJPE	PCKh	MPJPE	PA-MPJPE	PCKh
SPIN [24]	95.08	57.40	91.84	113.89	69.36	85.84	155.98	89.20	71.92
NeuralSMPL	**93.09**	**55.69**	**94.66**	**101.71**	**60.41**	**92.80**	**130.55**	**72.82**	**85.60**
HMR-EFT [19]	**89.88**	**53.51**	92.55	108.27	66.15	87.35	142.74	82.62	78.81
NeuralSMPL	90.31	55.13	**94.52**	**98.54**	**59.08**	**92.77**	**122.92**	**67.67**	**88.51**
PARE [22]	**81.10**	**50.77**	91.91	**93.26**	61.28	88.38	**117.59**	72.84	83.61
NeuralSMPL	89.91	54.75	**94.70**	97.95	**58.86**	**93.11**	121.88	**69.14**	**88.31**

Compare to Other Optimization-based Methods

Table 2: Comparison to other optimization-base methods. NeuralSMPL is the most robust to occlusion.

Method	3DPW			3DPW-AdvOcc@40			3DPW-AdvOcc@80		
	MPJPE	PA-MPJPE	PCKh	MPJPE	PA-MPJPE	PCKh	MPJPE	PA-MPJPE	PCKh
[19]	**89.88**	**53.51**	92.55	108.27	66.15	87.35	142.74	82.62	78.81
+ SMPLify [5]	105.15	63.74	93.16	112.88	68.23	89.02	143.89	79.37	82.92
+ EFT [19]	91.36	54.72	94.11	101.28	61.47	89.96	124.62	70.64	85.15
+ NeuralSMPL	90.31	55.13	**94.52**	**98.54**	**59.08**	**92.77**	**122.92**	**67.67**	**88.51**

Ablation Studies

Table 3: Ablation studies. All experiments are performed on 3DPW-AdvOcc@80 with the same initialization from HMR-EFT [19].

(a) Effectiveness of auxiliary loss functions.

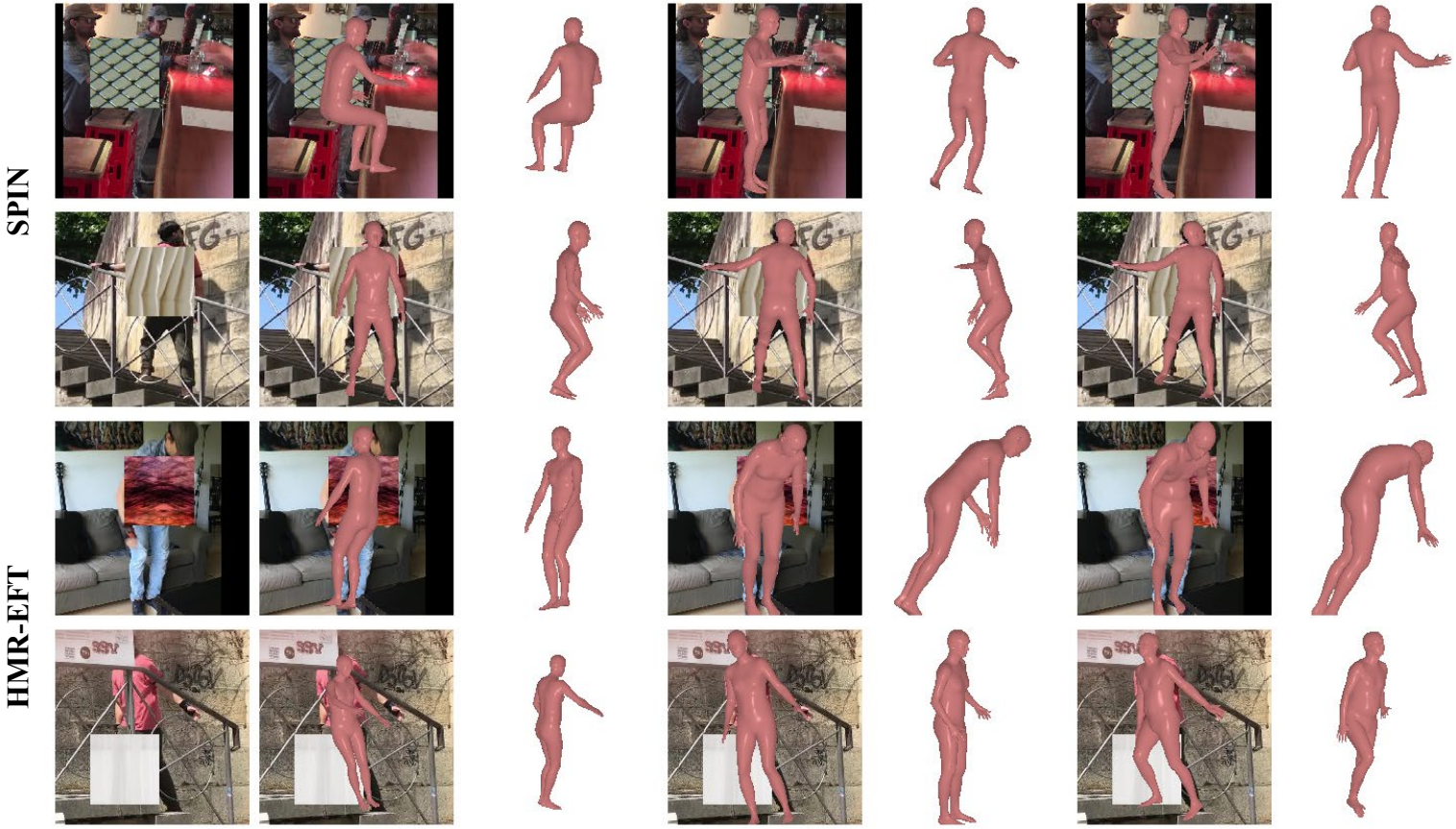
	NLL	Keyp. 2D	Part seg.	MPJPE	PA-MPJPE	PCKh
Initialization	-	-	-	142.74	82.62	78.81
NeuralSMPL		✓	✓	135.17	73.81	83.19
NeuralSMPL	✓			134.36	72.09	83.71
NeuralSMPL	✓		✓	130.46	71.40	84.63
NeuralSMPL	✓	✓		127.14	68.76	88.06
NeuralSMPL	✓	✓	✓	**122.92**	**67.67**	**88.51**

(b) Number of Gaussian mixtures components.

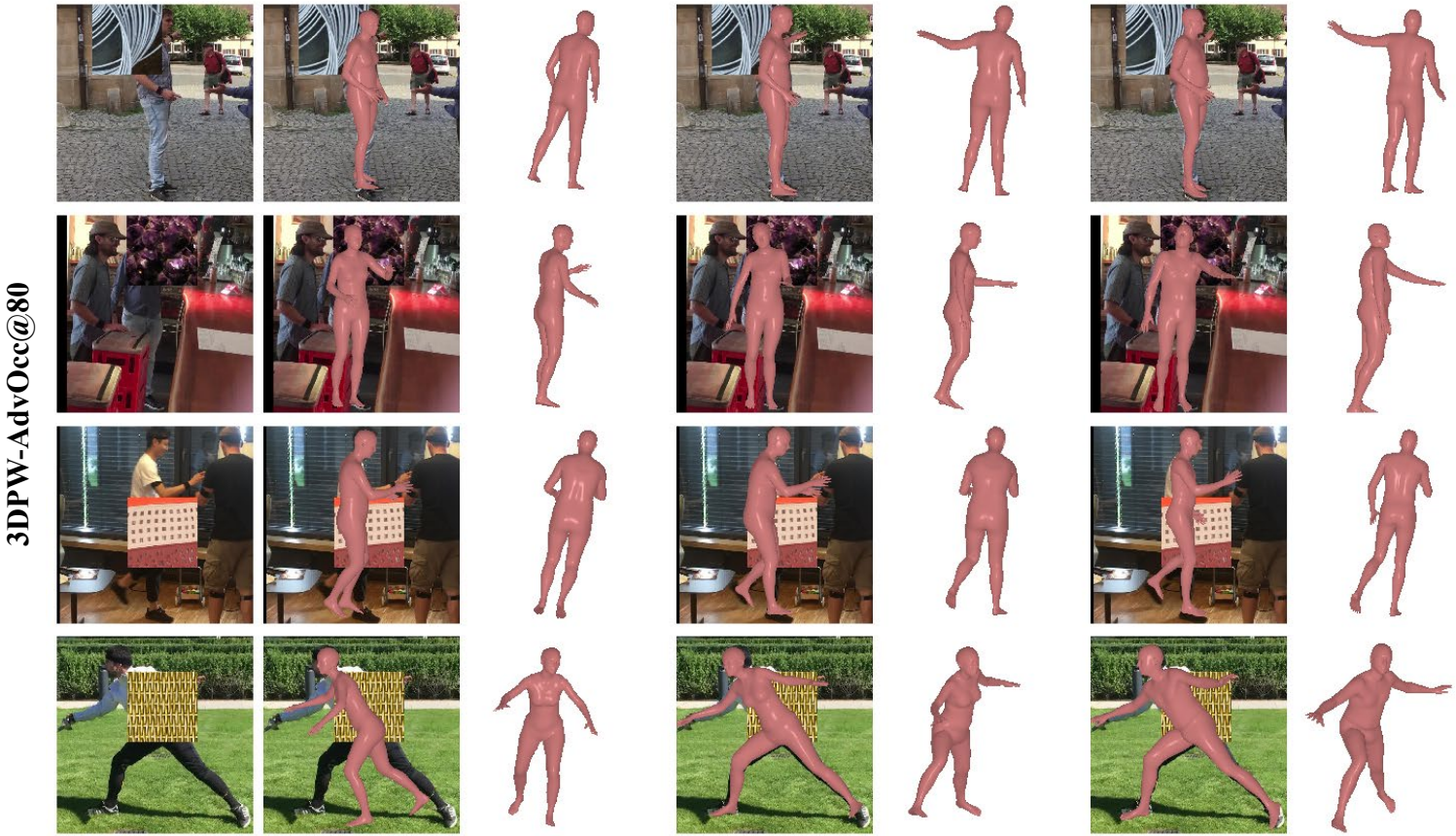
	K		
	1	3	5
MPJPE	127.09	122.92	**122.71**
PA-MPJPE	69.63	**67.67**	68.85
PCKh	**89.08**	88.51	87.87

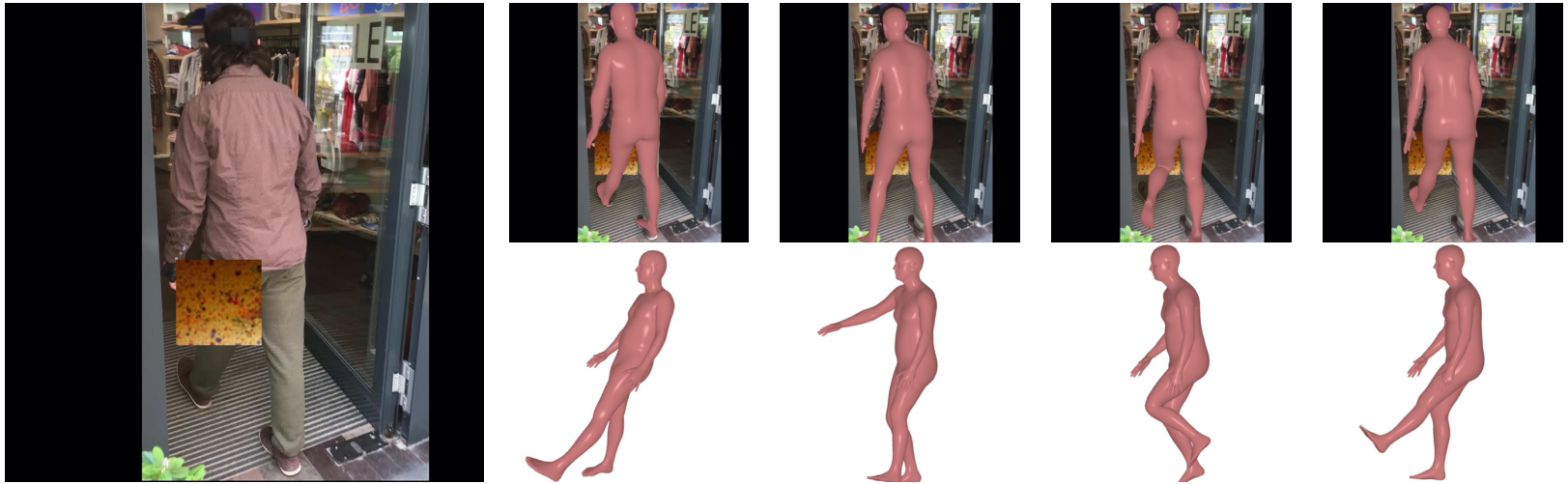
Table 1: Ablation on the softness of GMM components assignment.

	τ		
	5	10	100
MPJPE	**122.46**	122.92	125.02
PA-MPJPE	68.61	**67.67**	69.01
PCKh	**88.64**	88.51	88.62


SPIN
HMR-EFT
(a) Regressor
(b) NeuralSMPL
(c) GT


3DPW-AdvOcc@40
(a) PARE
(b) NeuralSMPL
(c) GT


3DPW-AdvOcc@80
(a) PARE
(b) NeuralSMPL
(c) GT

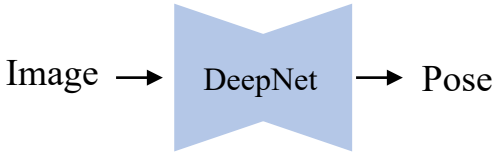


(a) Input Image (b) GT (c) HMR-EFT (d) HMR-EFT +SMPLify (e) HMR-EFT +NeuralSMPL

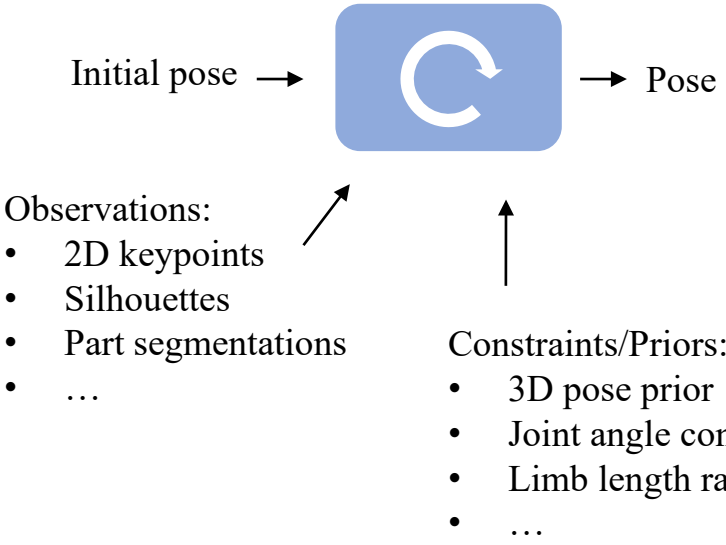
Questions?

Task: 3D Human Pose Estimation

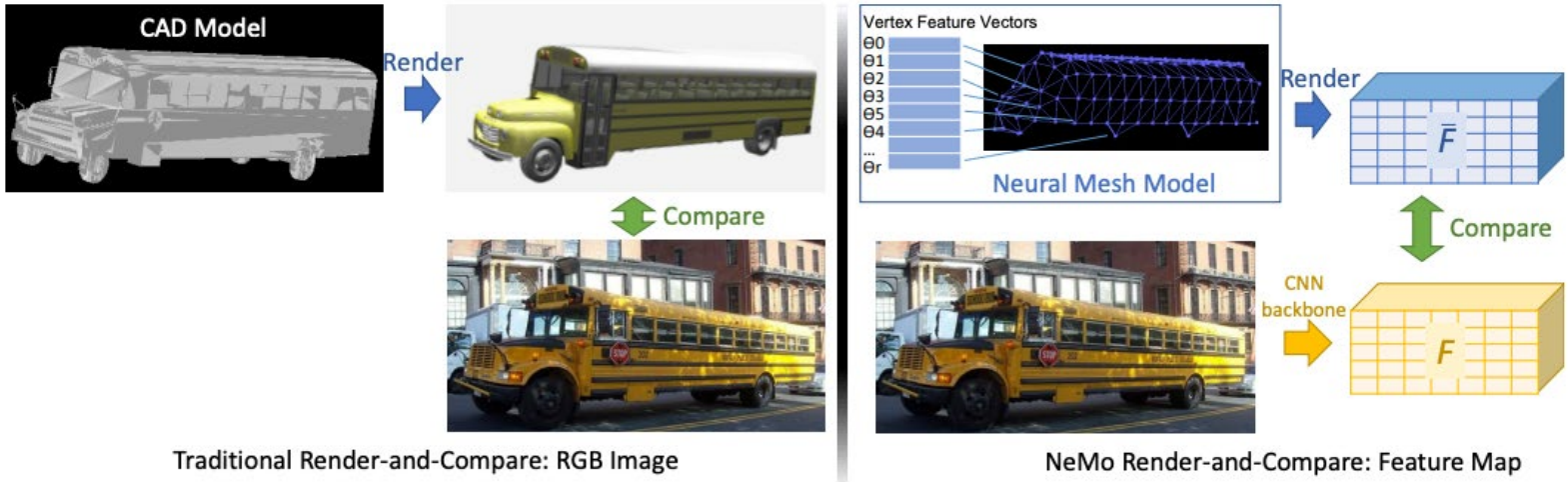
- Regression-based



- Optimization-based



Prior Work: Neural Mesh Models (NeMo)



- Generative model of feature vectors
  - Less sensitive to low-level appearance
  - Robust to occlusion