# CSC2457 3D & Geometric Deep Learning

## Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild

Shangzhe Wu, Christian Rupprecht, Andrea Vedaldi

Date: Tuesday, March 16, 2021

Presenter: Brendan Kolisnik

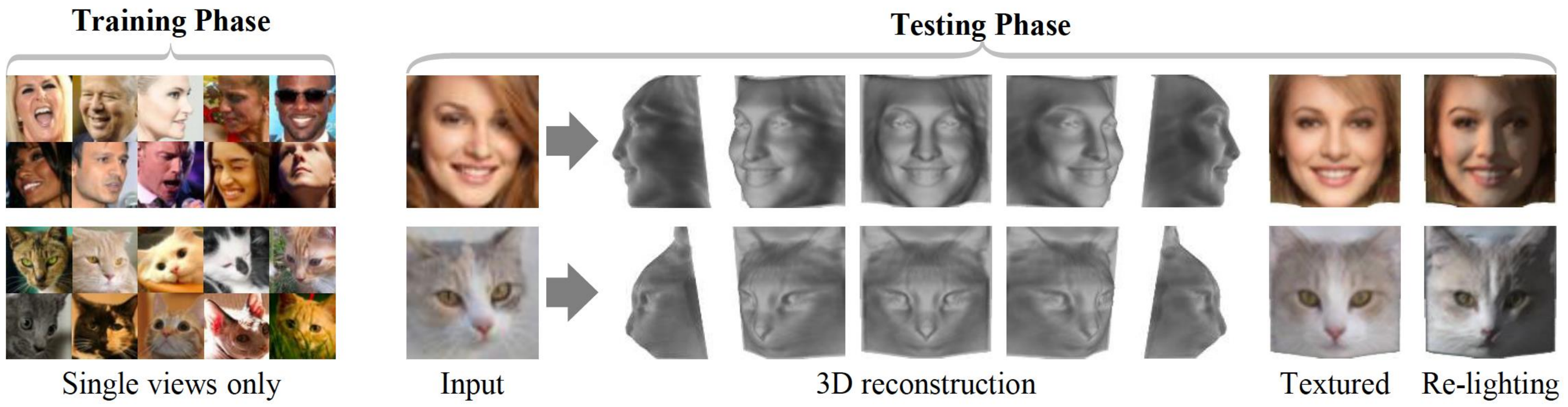
Instructor: Animesh Garg

UNIVERSITY OF TORONTO

# Motivation

- The majority of existing learning-based approaches to 3D reconstruction are supervised. The authors aim to solve 3D reconstruction from images under 2 major constraints.

1. No 2D or 3D ground truth information is available.
2. The model will only use single-view images, no multi-view inputs.



# Main Problem

- Performing 3D reconstruction from an image in an unsupervised setting is more usable than previous reconstruction efforts. Makes the algorithm more accessible to industry.
- One major challenge is that there is a low quantity of research for 3D reconstruction in an unsupervised setting. The authors are establishing the groundwork for this area.

# Prior work

| Paper | Supervision | Goals | Data |
|---|---|---|---|
| [47] | 3D scans | 3DMM | Face |
| [66] | 3DV, I | Prior on 3DV, predict from I | ShapeNet, Ikea |
| [1] | 3DP | Prior on 3DP | ShapeNet |
| [48] | 3DM | Prior on 3DM | Face |
| [17] | 3DMM, 2DKP, I | Refine 3DMM fit to I | Face |
| [15] | 3DMM, 2DKP, I | Fit 3DMM to I+2DKP | Face |
| [18] | 3DMM | Fit 3DMM to 3D scans | Face |
| [28] | 3DMM, 2DKP | Pred. 3DMM from I | Humans |
| [51] | 3DMM, 2DS+KP | Pred. N, A, L from I | Face |
| [64] | 3DMM, I | Pred. 3DM, VP, T, E from I | Face |
| [50] | 3DMM, 2DKP, I | Fit 3DMM to I | Face |
| [13] | 2DS | Prior on 3DV, pred. from I | Model/ScanNet |
| [30] | I, 2DS, VP | Prior on 3DV | ScanNet, PAS3D |
| [29] | I, 2DS+KP | Pred. 3DM, T, VP from I | Birds |
| [7] | I, 2DS | Pred. 3DM, T, L, VP from I | ShapeNet, Birds |
| [23] | I, 2DS | Pred. 3DV, VP from I | ShapeNet, others |
| [56] | I | Prior on 3DM, T, I | Face |
| [49] | I | Pred. 3DM, VP, $T^\dagger$ from I | Face |
| [22] | I | Pred. V, L, VP from I | ShapeNet |
| Ours | I | Pred. D, L, A, VP from I | Face, others |

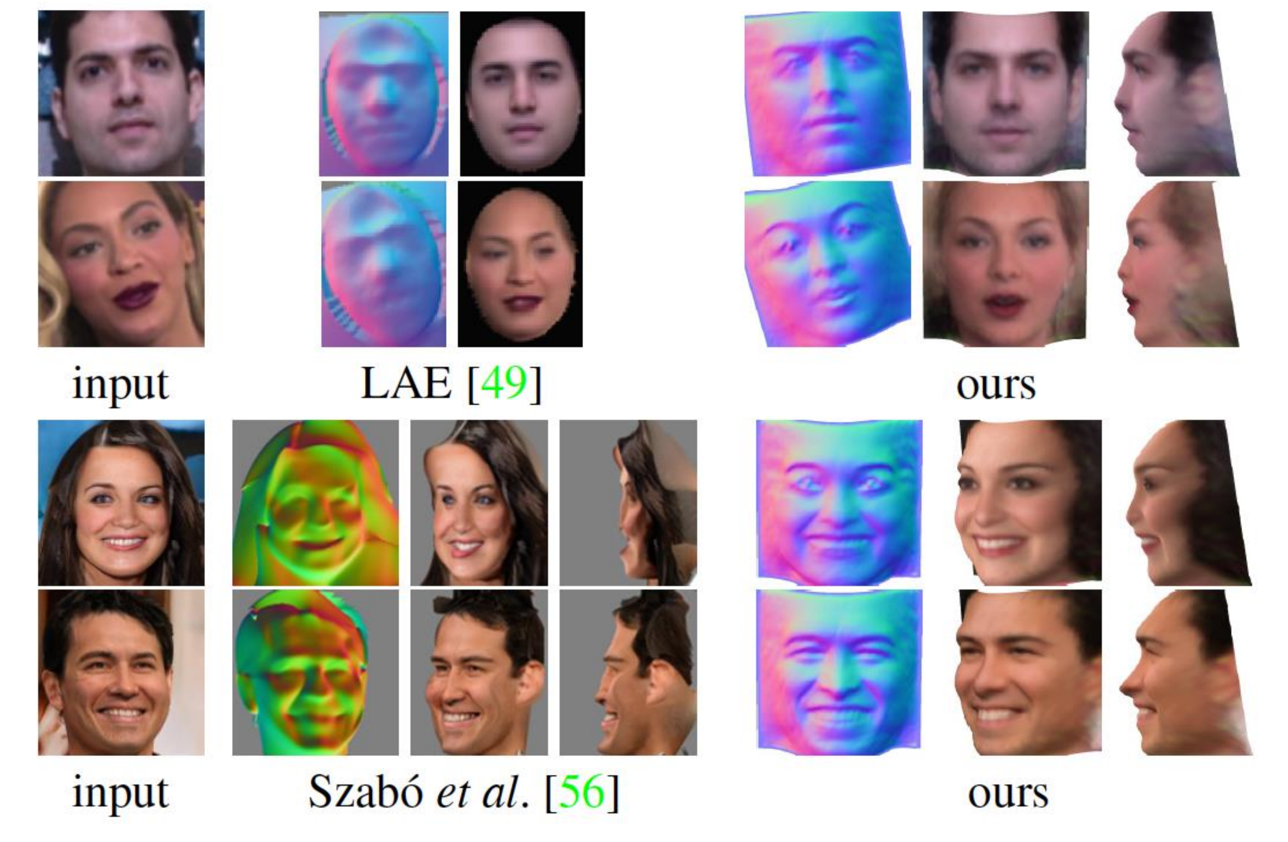
I: image, 3DMM: 3D morphable model, 2DKP: 2D keypoints, 2DS: 2D silhouette, 3DP: 3D points, VP: viewpoint, E: expression, 3DM: 3D mesh, 3DV: 3D volume, D: depth, N: normals, A: albedo, T: texture, L: light

# Contributions I

- The authors propose an unsupervised autoencoder approach to 3D reconstruction from images.
- AE factors each input image into depth, albedo, viewpoint and illumination without ground truth.
- This approach is ill-posed without additional constraints so the authors introduce additional assumptions such as taking advantage of bilateral symmetry in objects.
- One of the first works in unsupervised 3D reconstruction to show strong qualitative and quantitative results.

# Contributions II: Novelty

- The model takes into account that most objects are not totally symmetric by predicting a confidence of symmetry for each pixel.



# Problem Setting I

Image $\mathrm{I} : \Omega \to \mathbb{R}^3$

$$\Omega = \{0, \cdots, W-1\} \times \{0, \cdots, H-1\}$$

The goal is to learn a function, implemented as a neural network, that maps the image I to four factors. $(d, a, w, l)$ comprising a depth map $d : \Omega \to \mathbb{R}_+$, an albedo image $a : \Omega \to \mathbb{R}^3$, a global light direction $l \in \mathbb{S}^2$ and a viewpoint $w \in \mathbb{R}^6$ so that the image can be reconstructed from them.

$$\hat{\mathbf{I}} = \Pi(\Lambda(a, d, l), d, w)$$

Learning objective $\mathbf{I} \approx \hat{\mathbf{I}}$

# Problem Setting II

- Assume albedo and depth are symmetric about a fixed vertical plane.

$$[\text{flip a}]_{c,u,v} = a_{c,W-1-u,v}$$

$$\hat{\mathbf{I}}' = \Pi(\Lambda(a', d', l), d', w) \qquad a' = \text{flip } a \quad d' = \text{flip } d$$

Want: $\mathbf{I} \approx \hat{\mathbf{I}}$ and $\mathbf{I} \approx \hat{\mathbf{I}}'$

Predicted confidence map: $\sigma \in \mathbb{R}_+^{W \times H}$

# Approach

Only using this loss leads to blurry reconstructions

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) = -\frac{1}{|\Omega|} \sum_{uv \in \Omega} \ln \frac{1}{\sqrt{2}\sigma_{uv}} \exp -\frac{\sqrt{2}\ell_{1,uv}}{\sigma_{uv}}$$

where $\ell_{1,uv} = |\hat{\mathbf{I}}_{uv} - \mathbf{I}_{uv}|$

- Primary loss function is the negative log-likelihood of the factorized Laplacian distribution.

- To increase the visual fidelity the authors also compute an embedding for the two image reconstructions.

Kth layer of encoder predicts representation: $e^{(k)}(\mathbf{I}) \in \mathbb{R}^{C_k \times W_k \times H_k}$

# Approach: Loss Formulation

Perceptual Loss: $$\mathcal{L}_{\mathrm{p}}^{(k)}(\hat{\mathbf{I}}, \mathbf{I}, \sigma^{(k)}) = -\frac{1}{|\Omega_k|} \sum_{uv \in \Omega_k} \ln \frac{1}{\sqrt{2\pi(\sigma_{uv}^{(k)})^2}} \exp - \frac{(\ell_{uv}^{(k)})^2}{2(\sigma_{uv}^{(k)})^2}$$

where $\ell_{uv}^{(k)} = |e_{uv}^{(k)}(\hat{\mathbf{I}}) - e_{uv}^{(k)}(\mathbf{I})|$

Update loss definition: $$\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_p^{(k)}$$

Final loss definition: $$\mathcal{E}(\Phi; \mathbf{I}) = \mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) + \lambda_{\mathrm{f}} \mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}, \sigma')$$

where $\lambda_{\mathrm{f}} = 0.5$

# Method

$$\hat{\mathbf{I}} = \Pi(\Lambda(a, d, l), d, w)$$

With the 4 factored variables we can break this down into two steps.

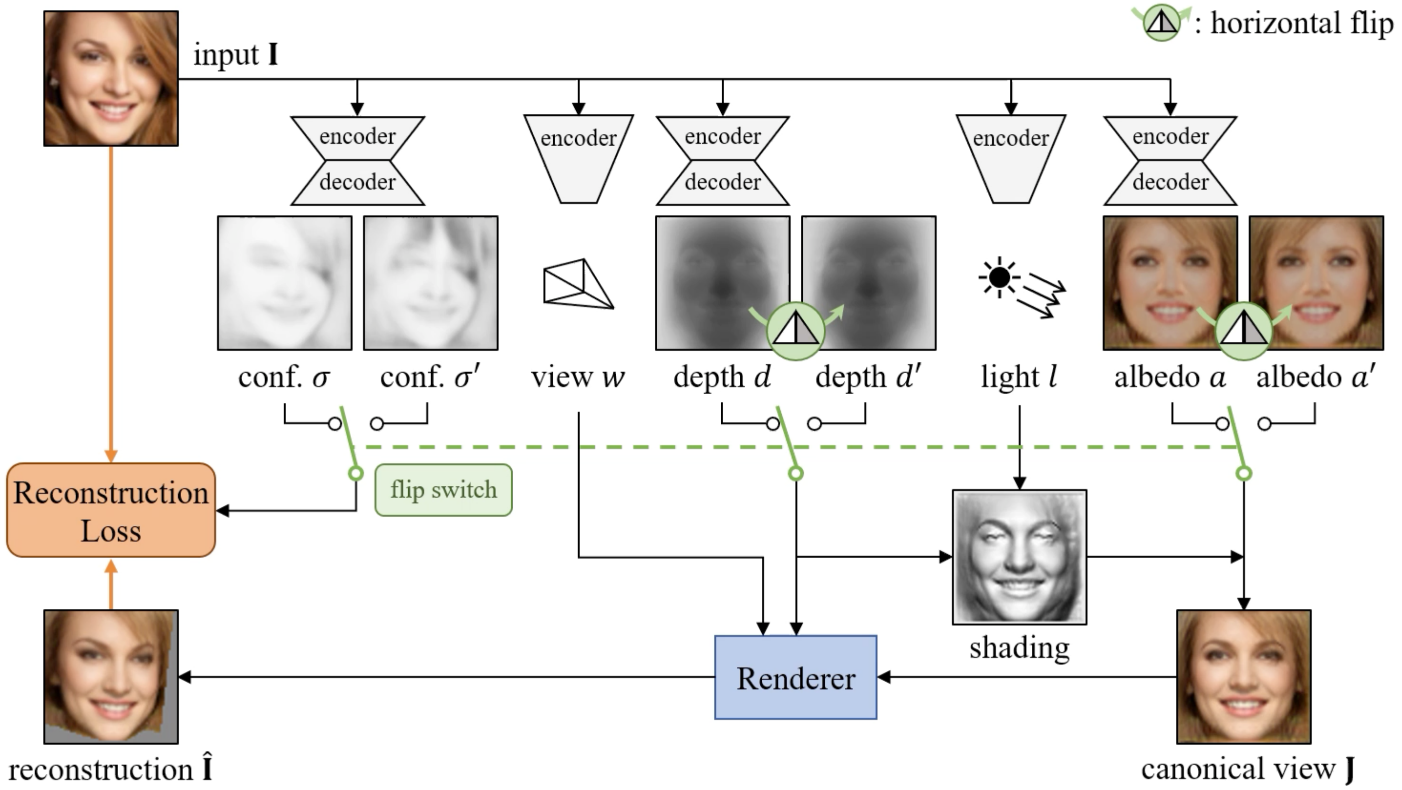
1. $\mathbf{J} = \Lambda(a, d, l)$

Calculate the canonical depth map with viewpoint w = 0

2. $\hat{\mathbf{I}} = \Pi(\mathbf{J}, d, w)$

Warp the canonical depth map and project to 2D to obtain the reconstructed image.

# Algorithm Overview




Two confidence-adjusted reconstruction losses are minimized at the same time with asymmetric weights.

# Experiment Metric

Scale Invariant Depth Error (SIDE): $$E_{\text{SIDE}}(\bar{d}, d^*) = \left(\frac{1}{WH}\sum_{uv}\Delta_{uv}^2 - \left(\frac{1}{WH}\sum_{uv}\Delta_{uv}\right)^2\right)^{\frac{1}{2}}$$

where $$\Delta_{uv} = \log \bar{d}_{uv} - \log d^*_{uv}$$

- SIDE measures the deviation from our predicted warped depth map to the ground-truth depth map.

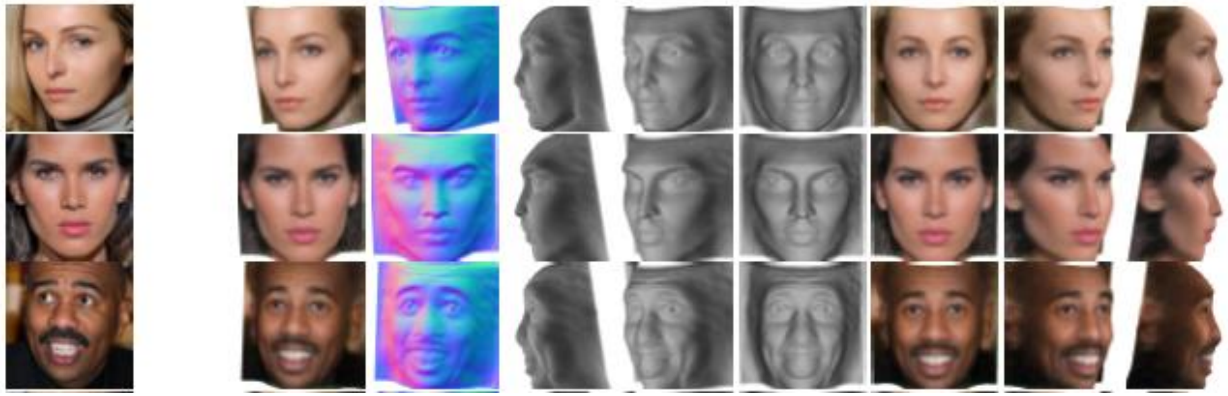
- Also look at mean angle deviation (MAD) between the normals computed from ground truth depth and predicted depth. MAD helps quantify how well surface details are captured

# Experiment Results

- Experiments performed using Basel Face Model synthetic generated face dataset (such that there is ground truth depth maps).
- Model approaches supervised performance.

| No | Baseline | SIDE ($\times 10^{-2}$) ↓ | MAD (deg.) ↓ |
|---|---|---|---|
| (1) | Supervised | 0.410 $\pm$0.103 | 10.78 $\pm$1.01 |
| (2) | Const. null depth | 2.723 $\pm$0.371 | 43.34 $\pm$2.25 |
| (3) | Average g.t. depth | 1.990 $\pm$0.556 | 23.26 $\pm$2.85 |
| (4) | Ours (unsupervised) | 0.793 $\pm$0.140 | 16.51 $\pm$1.56 |

Comparison with Baseline on BFM



Unsupervised Reconstruction

All models trained for 50k iterations.

# Experiment Results: Ablation Study Visualized

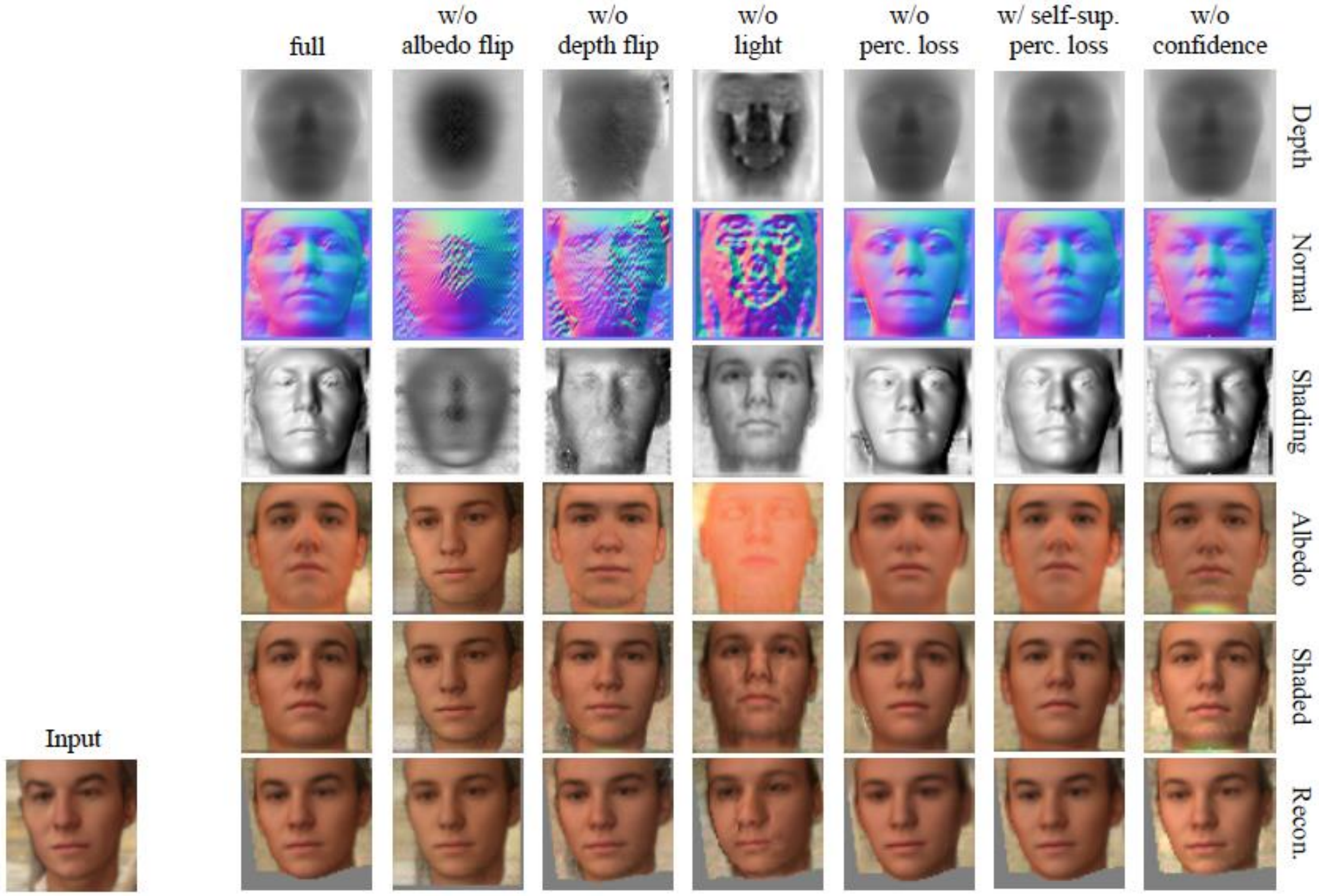


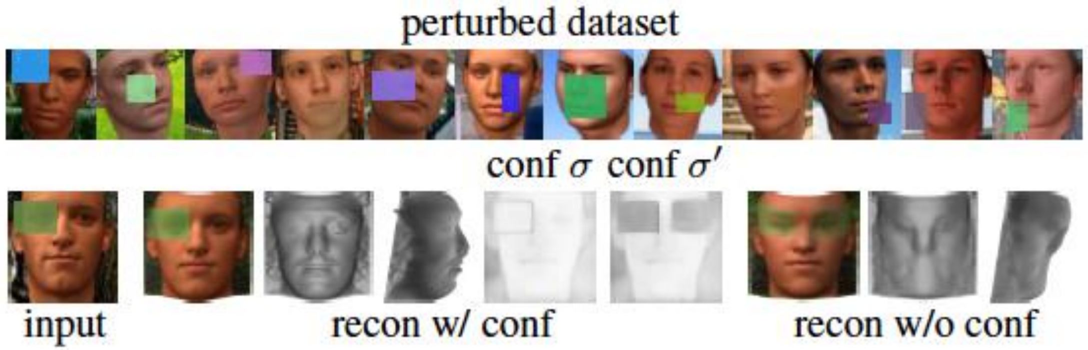

Figure 9: **Qualitative results of the ablated models.**

| No | Method | SIDE ($\times 10^{-2}$) $\downarrow$ | MAD (deg.) $\downarrow$ |
|---|---|---|---|
| (1) | Ours full | 0.793 $\pm$0.140 | 16.51 $\pm$1.56 |
| (2) | w/o albedo flip | 2.916 $\pm$0.300 | 39.04 $\pm$1.80 |
| (3) | w/o depth flip | 1.139 $\pm$0.244 | 27.06 $\pm$2.33 |
| (4) | w/o light | 2.406 $\pm$0.676 | 41.64 $\pm$8.48 |
| (5) | w/o perc. loss | 0.931 $\pm$0.269 | 17.90 $\pm$2.31 |
| (6) | w/ self-sup. perc. loss | 0.815 $\pm$0.145 | 15.88 $\pm$1.57 |
| (7) | w/o confidence | 0.829 $\pm$0.213 | 16.39 $\pm$2.12 |

Ablation study of all model features

# Experiment Results: Perturbation Tests

- On the ablation study the SIDE and MAD are good even without confidence but keep in mind that BFM is a face dataset with lots of symmetry.
- Authors show that confidence is necessary for images with lots of asymmetry.



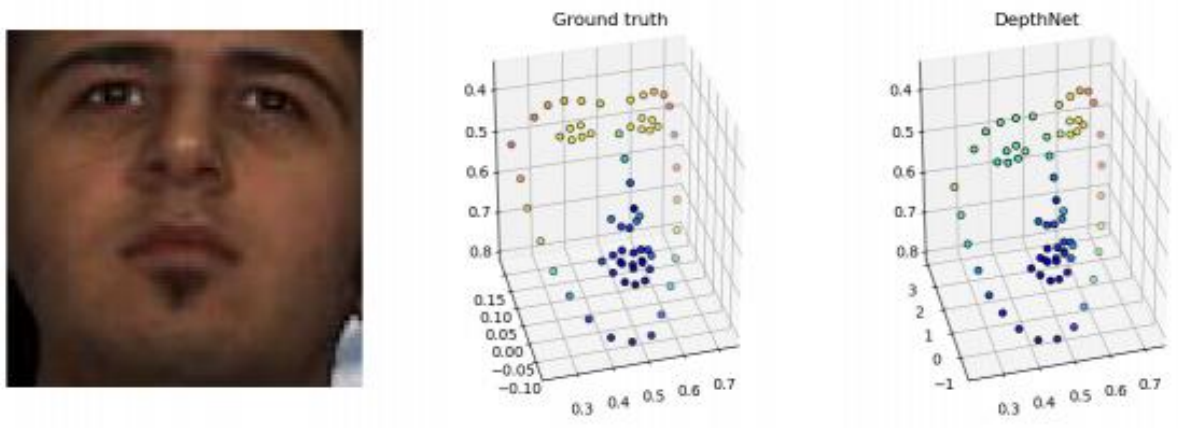

| | SIDE ($\times 10^{-2}$) $\downarrow$ | MAD (deg.) $\downarrow$ |
|---|---|---|
| No perturb, no conf. | 0.829 $\pm 0.213$ | 16.39 $\pm 2.12$ |
| No perturb, conf. | 0.793 $\pm 0.140$ | 16.51 $\pm 1.56$ |
| Perturb, no conf. | 2.141 $\pm 0.842$ | 26.61 $\pm 5.39$ |
| Perturb, conf. | 0.878 $\pm 0.169$ | 17.14 $\pm 1.90$ |

Perturbation tests with and without confidence

# Additional Quantitative Results

| | Depth Corr. ↑ |
|---|---|
| Ground truth | 66 |
| AIGN [61] (**supervised**, from [40]) | 50.81 |
| DepthNetGAN [40] (**supervised**, from [40]) | 58.68 |
| MOFA [57] (**model-based**, from [40]) | 15.97 |
| DepthNet [40] (from [40]) | 26.32 |
| DepthNet [40] (from GitHub) | 35.77 |
| Ours | 48.98 |
| Ours (w/ CelebA pre-training) | 54.65 |

Table 5: **3DFAW keypoint depth evaluation.** Depth correlation between ground truth and prediction evaluated at 66 facial keypoint locations.



# Discussion of Results

- Competitive with supervised models on face datasets.

- Qualitatively the model is much better than previous unsupervised works.

- Authors have shown that symmetry and illumination are strong cues for shape and aid the model in predictive ability.

# Critique / Limitations

- The authors acknowledge the model has limitations due to architecture design.

a: extreme lighting b: noisy texture c: extreme pose

- Authors should provide more information on confidence maps since it is one of the more novel contributions for modelling asymmetry.
- Additionally, the model does not output a full 3D mesh but depth map with additional info.
- Other works such as Unsupervised Learning of Category-Specific Symmetric 3D Keypoints from Point Sets show that due to symmetry assumptions it only works for a single face.

# Contributions (Recap)

- Authors have introduced a new model for 3D reconstruction from images that is unsupervised.
- Prior works had been supervised with ground truth meshes, silhouettes etc.
- This work can exceed supervised performance.
- The model uses encoder-decoder networks to extract depth, albedo, viewpoint and illumination.
- For asymmetrical inputs the confidence of symmetry was key.