Toward Realistic Single-View 3D Object Reconstruction with Unsupervised Learning from Multiple Images

Long-Nhat Ho, Anh Tuan Tran, Quynh Phung, Minh Hoai

VinAI Research, VinUniversity, Stony Brook University
Motivation
Problem

Recover 3D structure (shape + texture) of an object of a known category in a single image

✘ Ill-posed problem
✔ Human is very good at this task via learning 3D shape prior
Problem

How to learn the 3D shape prior?

- **Supervised**
  - Require massive 3D data → hard to acquire

- **Unsupervised**
  - Observe 2D images of the same category
Previous approach – LeSym*

Only symmetric objects !!!

*S. Wu, C. Rupprecht, and A. Vedaldi. “Unsupervised learning of probably symmetric deformable 3d objects from images in the wild”. In CVPR 2020.
Our solution?

- Many datasets have *multiple images* for each *object instance*.
  - Cover symmetric objects

- Shape consistency
Learning from Multi-Image Datasets - LeMul
LeMul System
Recon.
image 1
Recon.
image 2

Input 1
Input 2

Decomposing
Network

Canon. Albedo1
Canon. Depth1

Canon. Albedo2
Canon. Depth2

View 1
Light 1

View 2
Light 2

Recon. image 1
Recon. image 2

Reconstruction loss

Cross-view consistency loss

*Note that we omit the confidence maps in this figure for simplicity
LeMul system

\[ \mathbb{L}^{al}(I, a, d) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left\| \sum_{p_k \in N(p)} w_k^c w_k^d (a(p) - a(p_k)) \right\|^2 \]

Where:

- \( N(p) \): the neighbors of a pixel
- \( w_k^c \): the intensity weighting term
- \( w_k^d \): the depth weighting term
Results
Qualitative results

| Input | LeSym | Ours |
|-------|-------|------|
| BFM   | ![BFM Images] | ![BFM Images] |
| CelebA| ![CelebA Images] | ![CelebA Images] |
| CatFaces | ![Cat Faces Images] | ![Cat Faces Images] |

single-image, symmetric objects
### Qualitative results

|       | Input          | LeSym          | Ours          |
|-------|----------------|----------------|---------------|
| BFM   | ![BFM Image]   | ![LeSym Image] | ![Ours Image] |
| CelebA| ![CelebA Image]| ![LeSym Image] | ![Ours Image] |
| CatFaces | ![CatFaces Image] | ![LeSym Image] | ![Ours Image] |
| MultiPIE | ![MultiPIE Image] | ![LeSym Image] | ![Ours Image] |

**multi-view dataset**
### Qualitative results

| Dataset   | Input | LeSym | Ours |
|-----------|-------|-------|------|
| BFM       | ![BFM](image) | ![LeSym BFM](image) | ![Ours BFM](image) |
| CelebA    | ![CelebA](image) | ![LeSym CelebA](image) | ![Ours CelebA](image) |
| CatFaces  | ![CatFaces](image) | ![LeSym CatFaces](image) | ![Ours CatFaces](image) |
| MultiPIE  | ![MultiPIE](image) | ![LeSym MultiPIE](image) | ![Ours MultiPIE](image) |
|CASIA      | ![CASIA](image) | ![LeSym CASIA](image) | ![Ours CASIA](image) |

*image collection dataset*
## Qualitative results

| Input          | LeSym                        | Ours                        |
|----------------|------------------------------|-----------------------------|
| BFM            | ![BFM Image](image1)         | ![Ours Image](image2)       |
| CelebA         | ![LeSym Image](image3)       | ![Ours Image](image4)       |
| CatFaces       | ![CatFaces-bfm](image5)      | ![CatFaces-ours](image6)    |
| MultiPIE       | ![MultiPIE-bfm](image7)      | ![MultiPIE-ours](image8)    |
| CASIA          | ![CASIA-bfm](image9)         | ![CASIA-ours](image10)      |
| YTF            | ![YTF-bfm](image11)          | ![YTF-ours](image12)        |

*video dataset*
Cat Faces (single + symmetric)
Multi-PIE (multi-view)
CASIA-WebFace (image collection)
CASIA-WebFace (image collection)

In-the-wild

Input
LeSym (CelebA)
LeMul (CASIA)
Quantitative results

- Better surface reconstruction on BFM
- Better voted via user surveys on all datasets

| No | Baseline                  | SIDE(x10^{-2})↓ | MAD(deg.)↓ |
|----|--------------------------|-----------------|------------|
| (1) | Supervised               | 0.410 ± 0.103   | 10.78 ± 1.01 |
| (2) | Const. null depth        | 2.723 ± 0.371   | 43.34 ± 2.25 |
| (3) | Average G.T. depth       | 1.990 ± 0.556   | 23.26 ± 2.85 |
| (4) | LeSym                    | 0.793±0.140     | 16.51 ± 1.56 |
| (5) | LeMul (proposed)         | 0.834 ± 0.169   | 15.49±1.50  |

BFM results comparison with baselines.
THANK YOU
https://github.com/VinAIResearch/LeMul