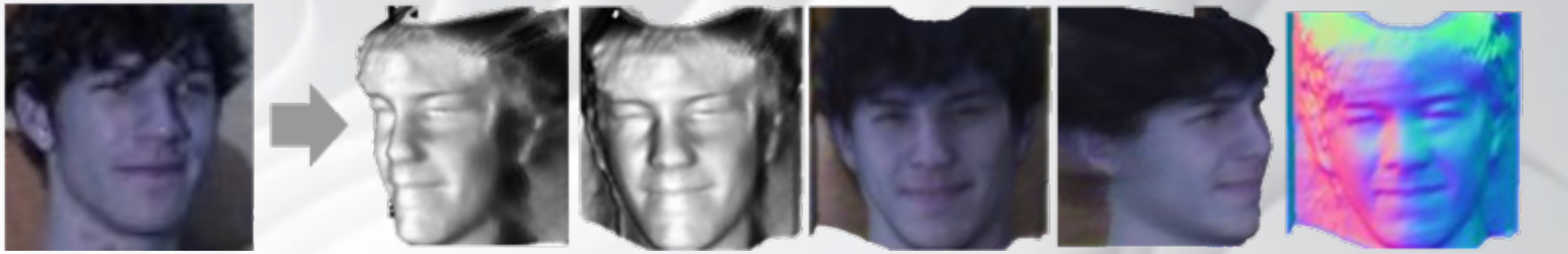
2021 ICCV OCTOBER 11-17 VIRTUAL



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

Long-Nhat Ho Anh Tuan Tran[1,2] Quynh Phung[1] Minh Hoai[1,3]

VinAI Research[1] VinUniversity[2] Stony Brook Univeristy[3]

VinAI RESEARCH

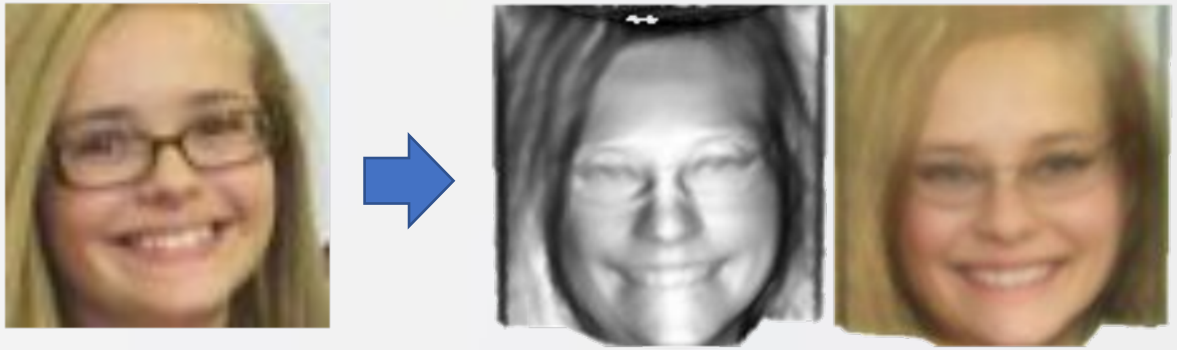
VINUNIVERSITY

Stony Brook University

# Motivation

# Problem

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

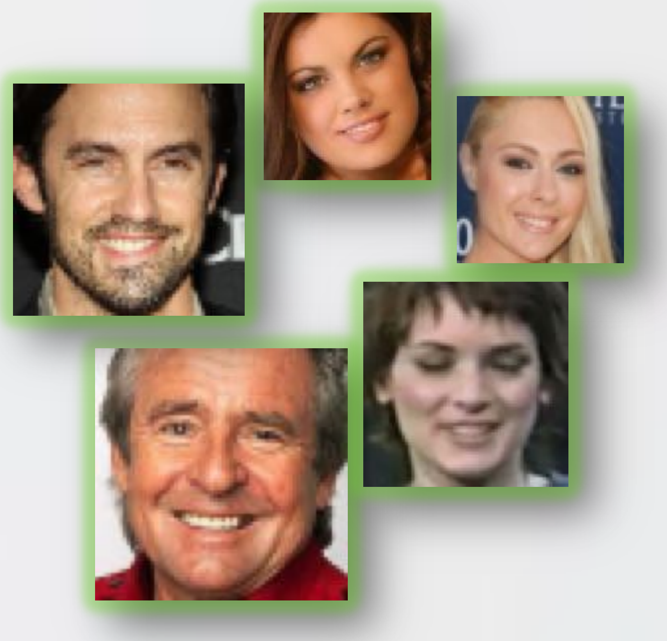


- ✗ Ill-posed problem
- ✓ Human is very good at this task via learning **3D shape prior**

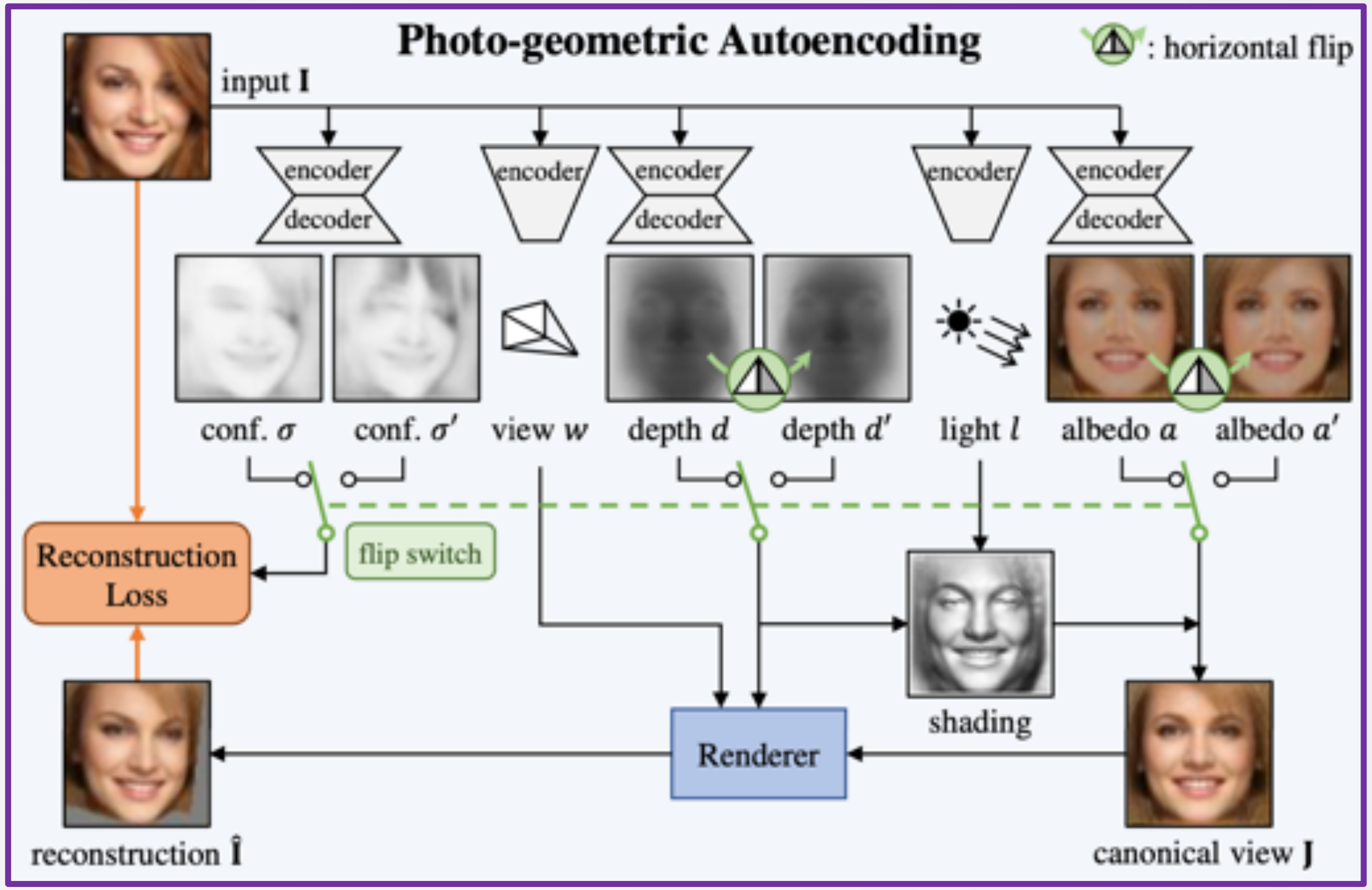
# Problem

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 !!!

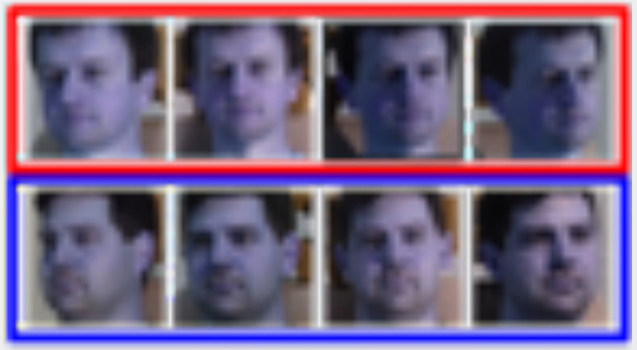
* S. Wu, C. Rupprecht, and A. Vedaldi. "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild". In *CVPR 2020.*

# Our solution?

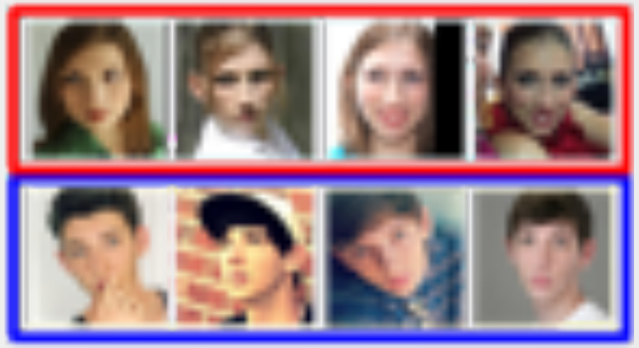
- Many datasets have *multiple images* for each *object instance*
    - ✓ Cover symmetric objects

- Shape consistency

Multi-view

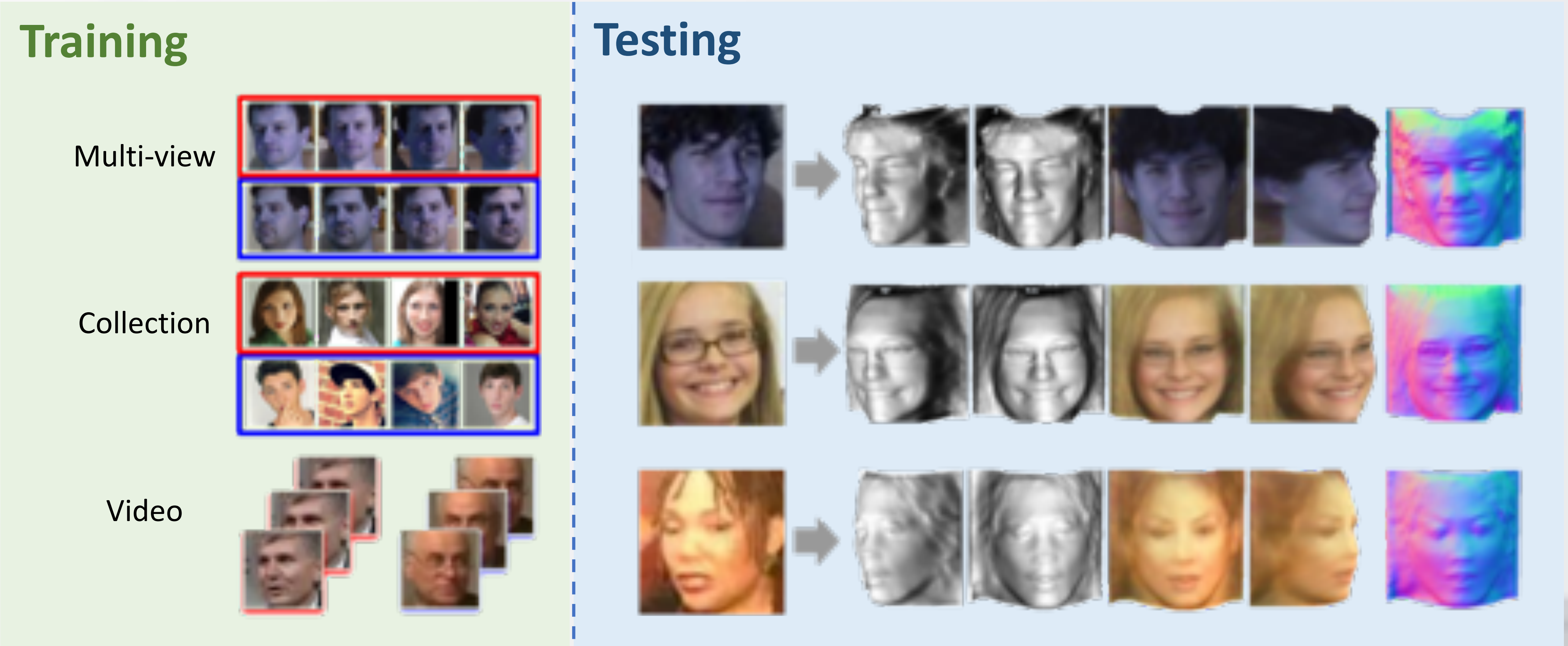


Collection



Video

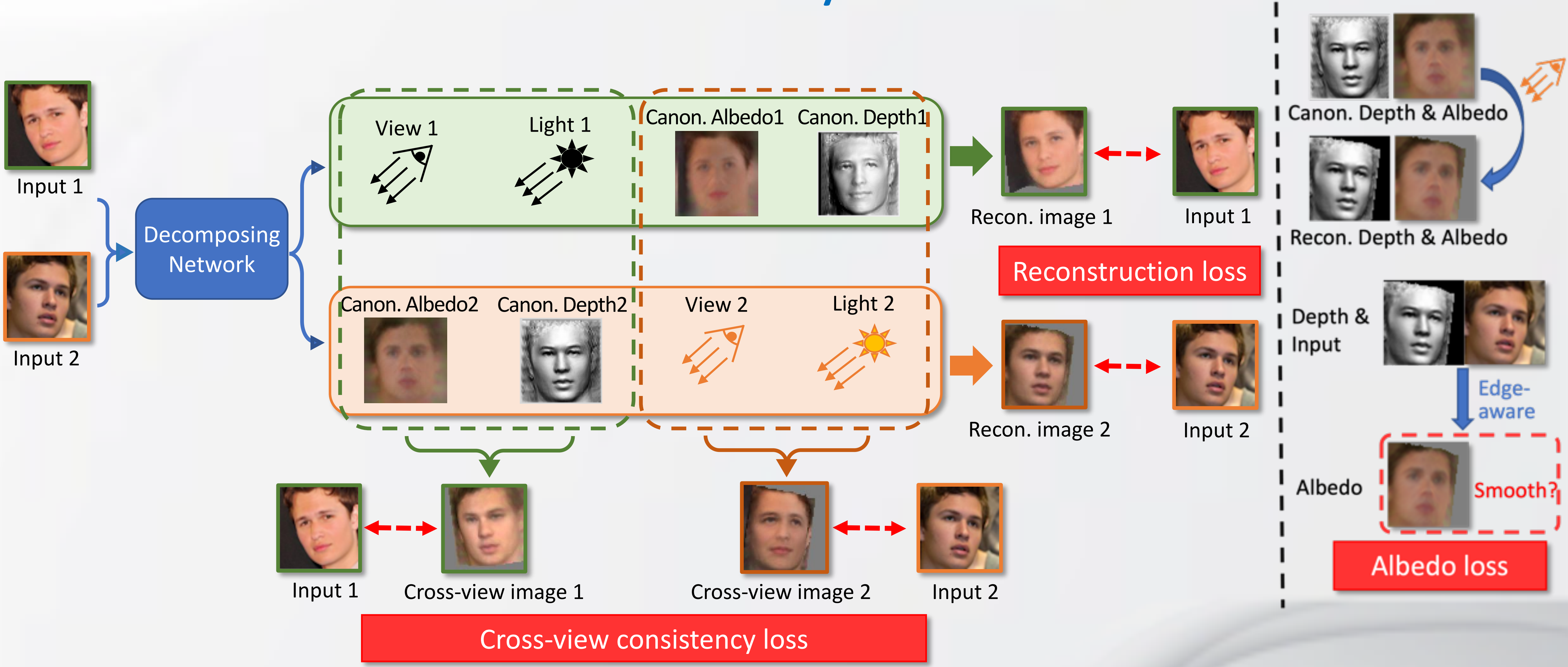
# LeMul



**Le**arning from **Mu**lti-Image Datasets - ***LeMul***

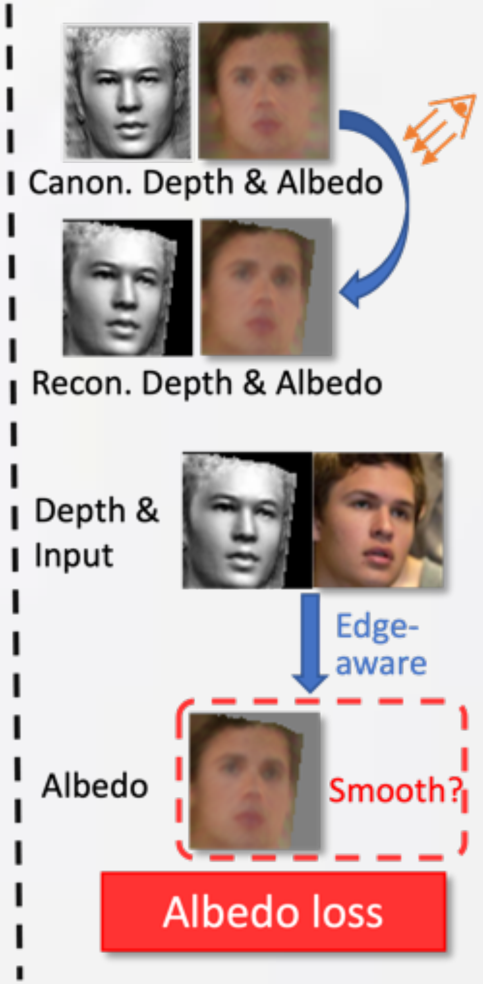
# *LeMul* System

# LeMul system



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

# LeMul system



$$\mathbb{L}^{al}(\mathbf{I}, a, d) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \Big\| \sum_{p_k \in \mathcal{N}(p)} w_k^c w_k^d (a(p) - a(p_k)) \Big\|^2$$

Where:

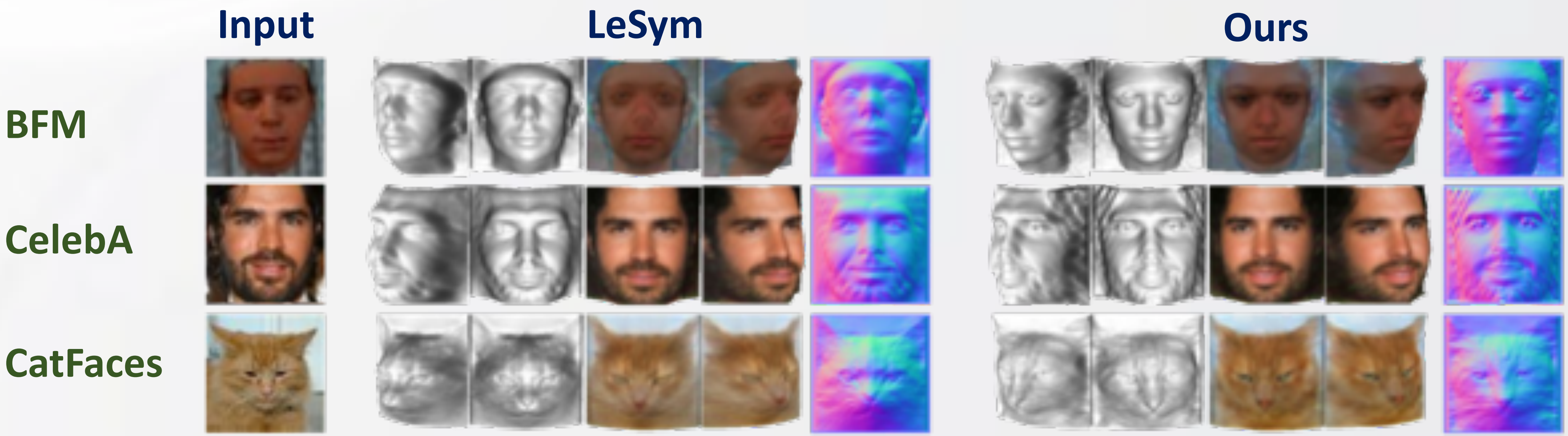
N(p) : the neighbors of a pixel

$w_k^c$: the intensity weighting term

$w_k^d$: the depth weighting term

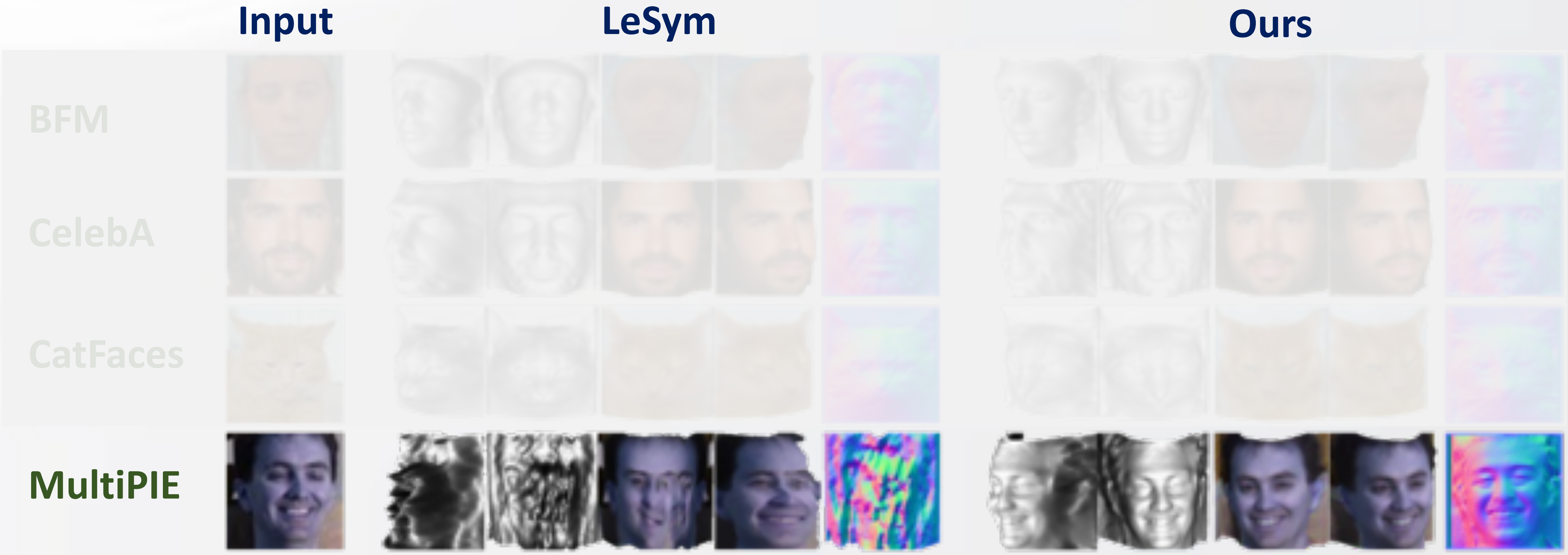
# Results

# Qualitative results



single-image, symmetric objects

# Qualitative results



multi-view dataset

# Qualitative results

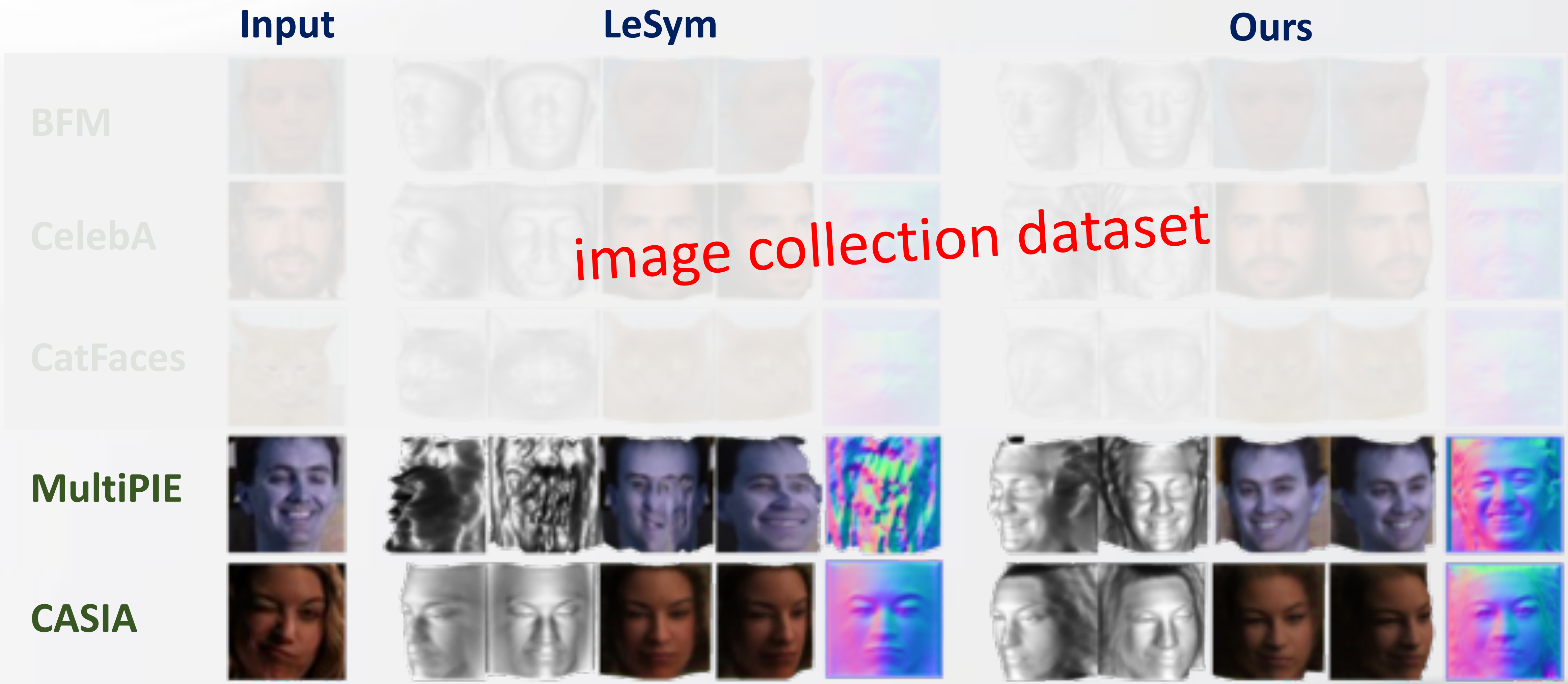
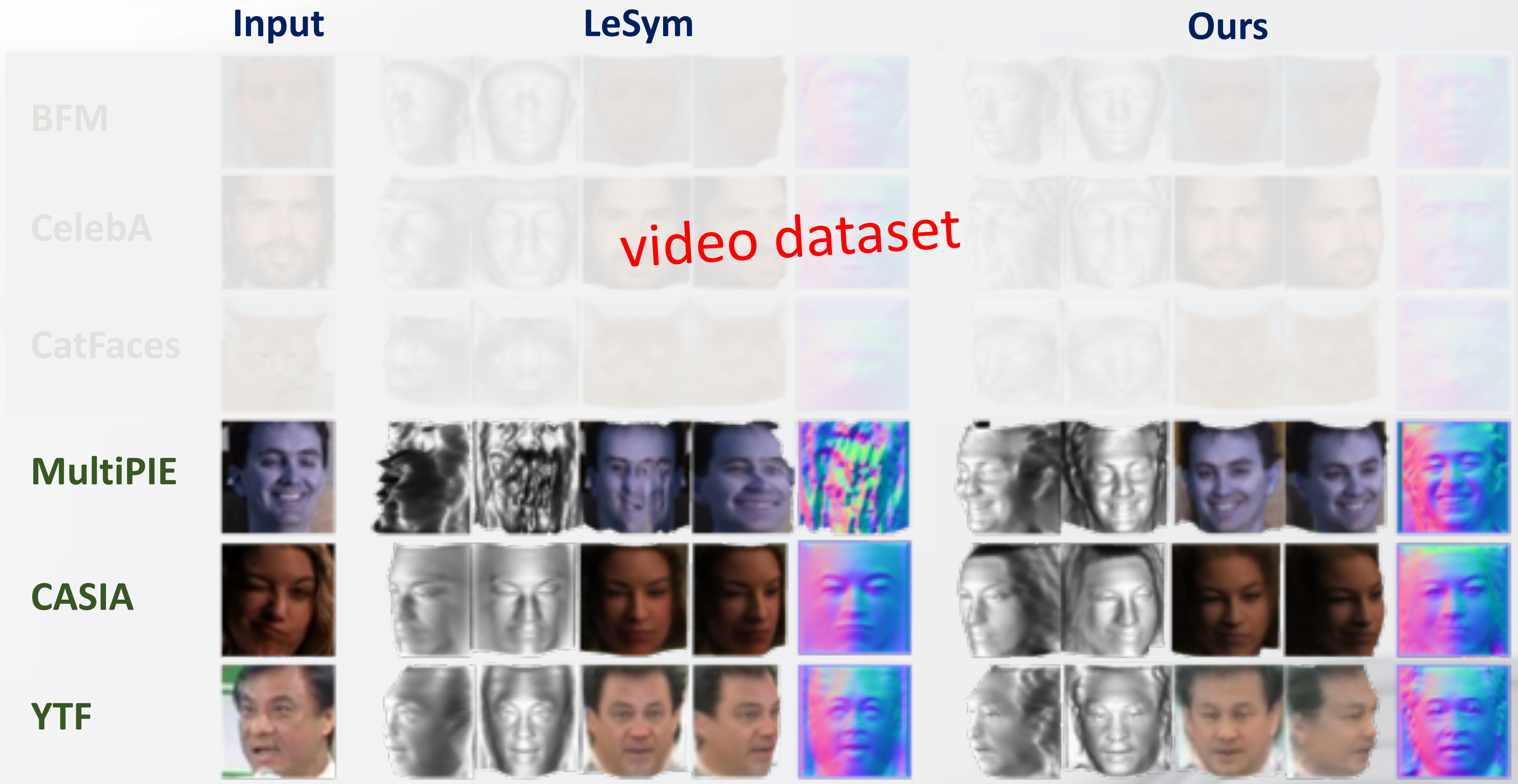


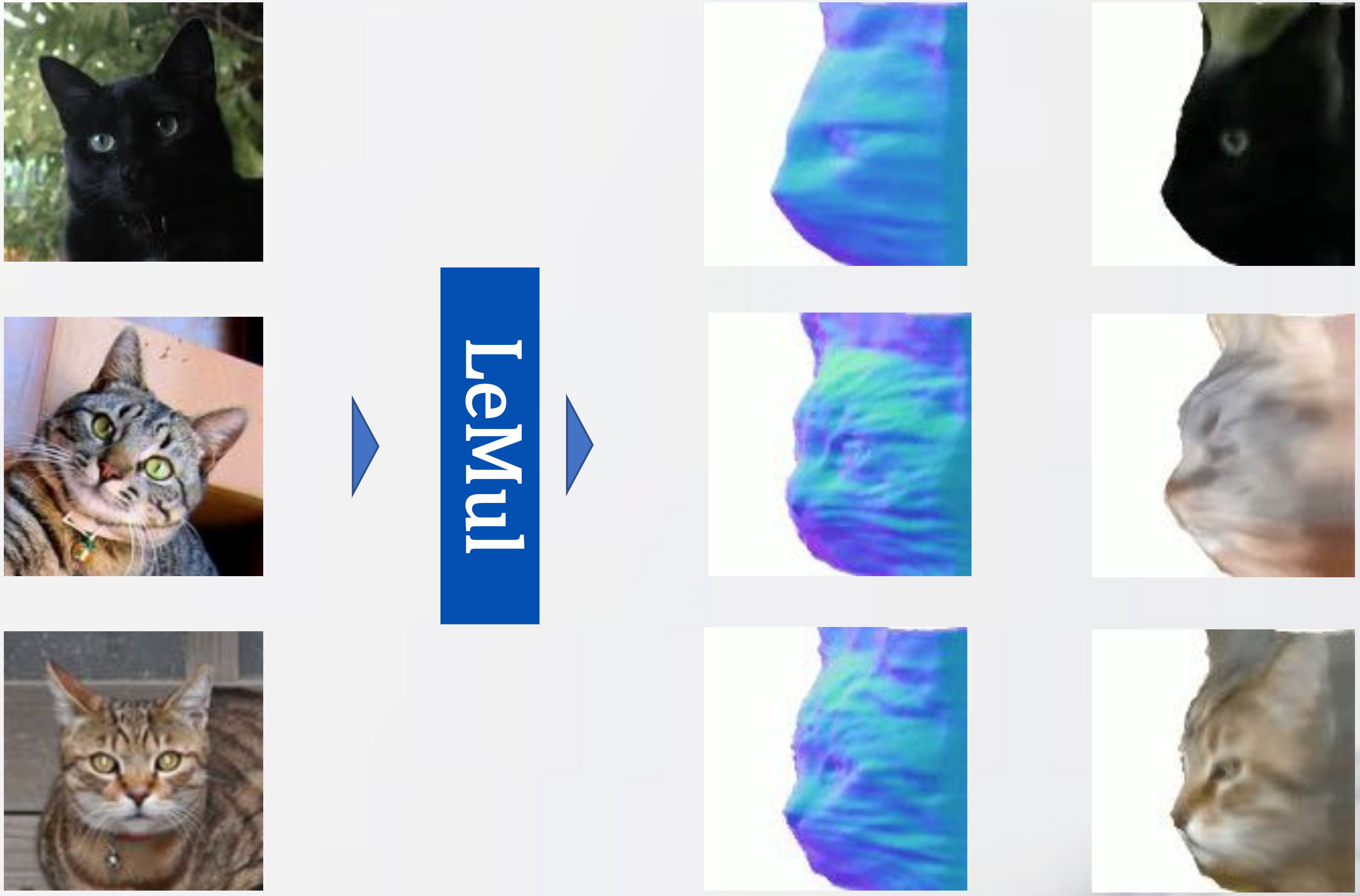
image collection dataset

# Qualitative results

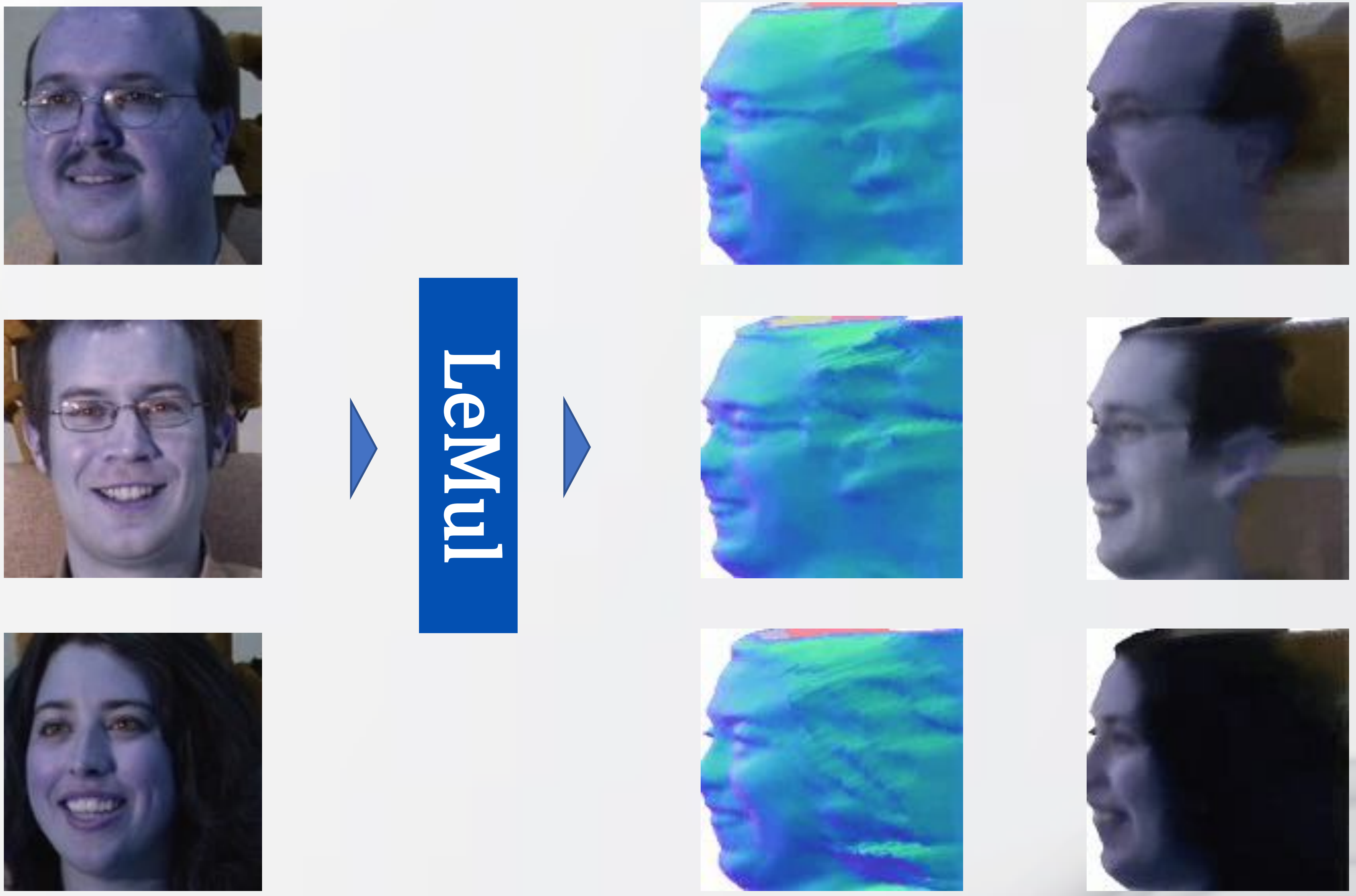


video dataset

# Cat Faces (single + symmetric)

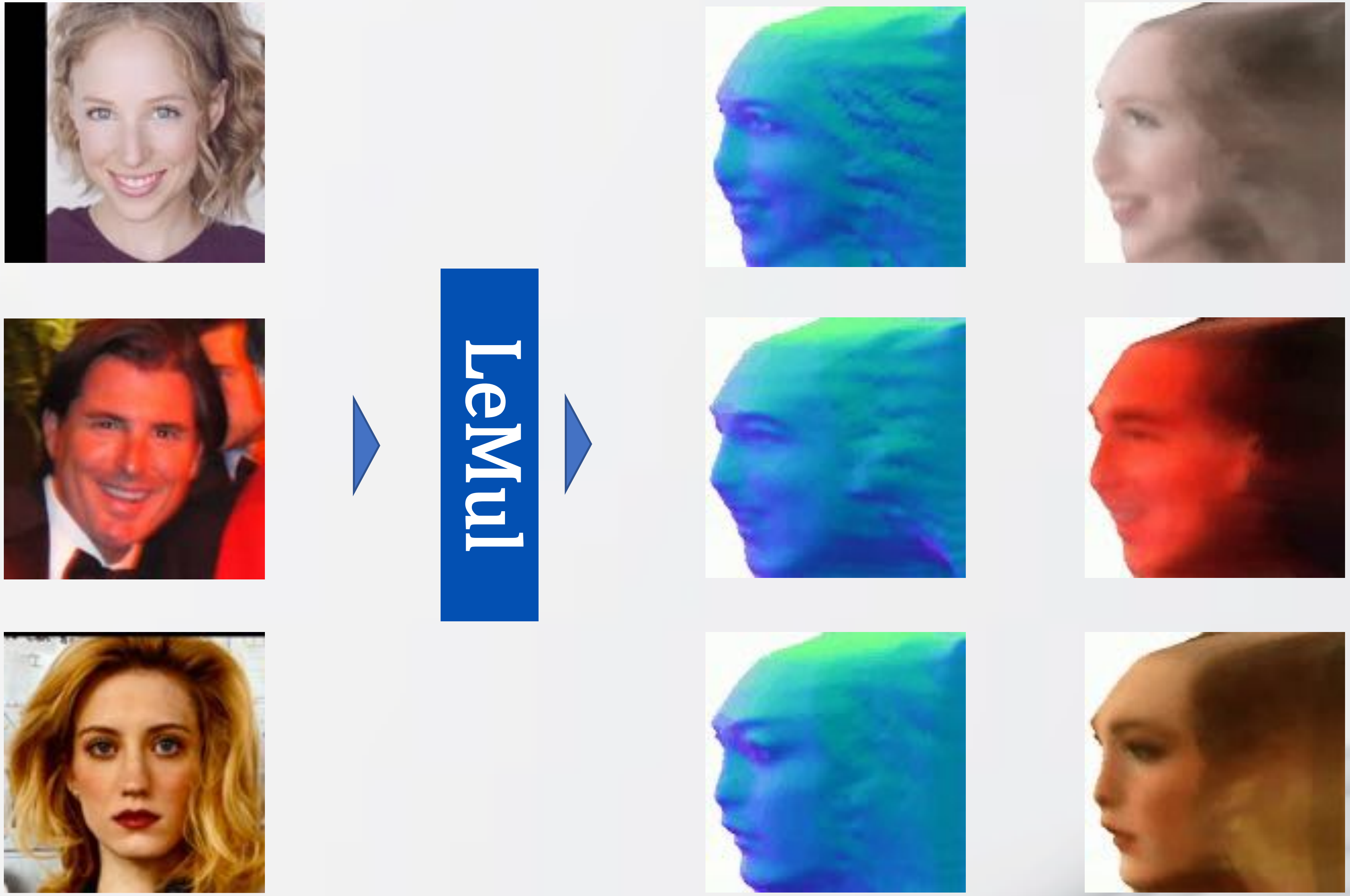


# Multi-PIE (multi-view)



LeMul

# CASIA-WebFace (image collection)



LeMul

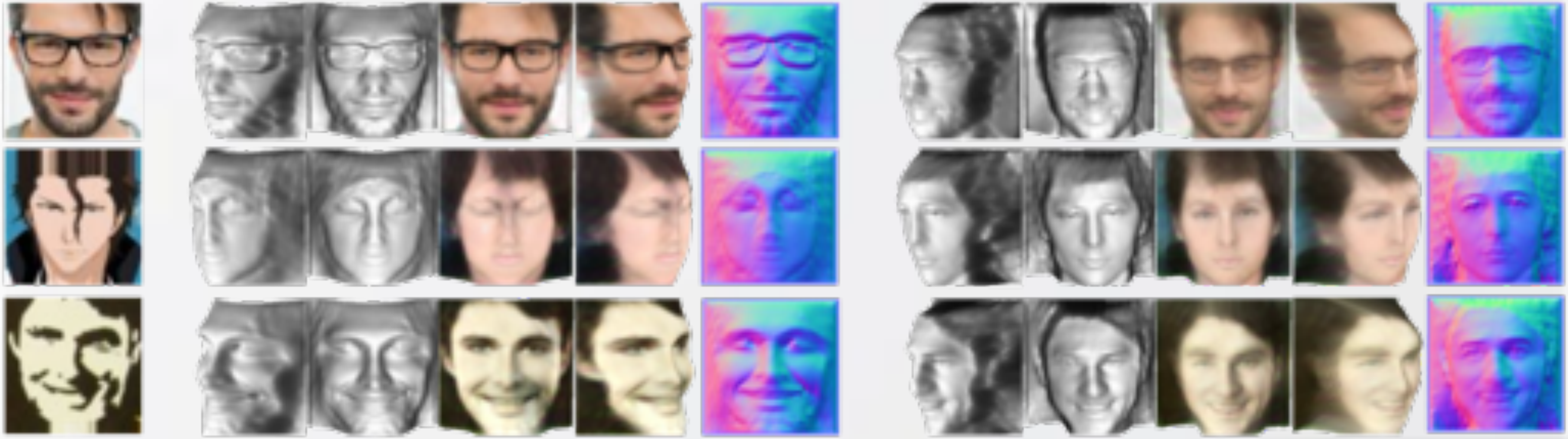
# CASIA-WebFace (image collection)

In-the-wild

**Input**

**LeSym (CelebA)**

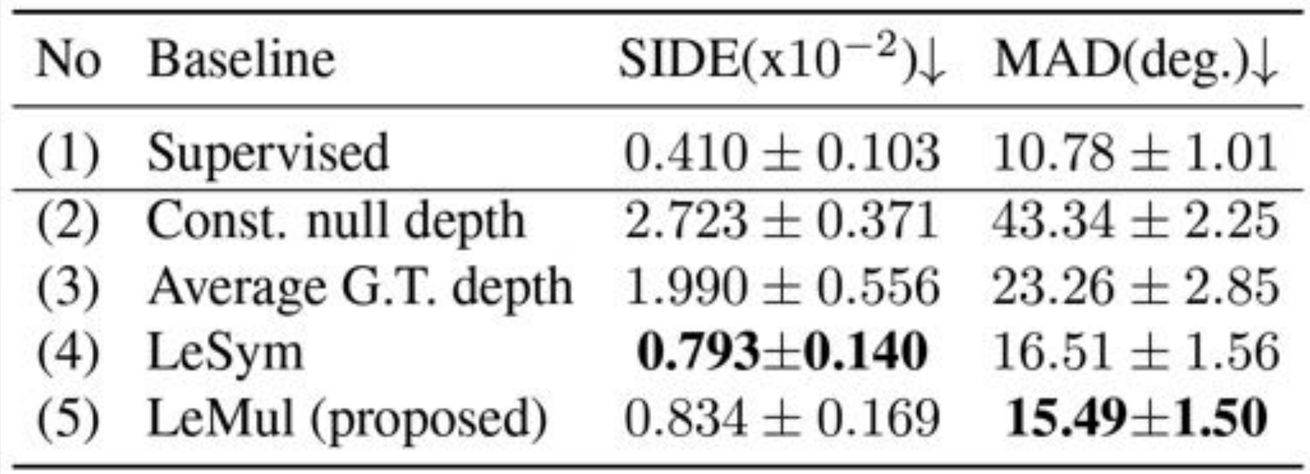
**LeMul (CASIA)**



# Quantitative results

✓Better surface reconstruction on BFM

✓Better voted via user surveys on all datasets



| No | Baseline | SIDE(x$10^{-2}$)↓ | 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) | LeSym | $\mathbf{0.793 \pm 0.140}$ | $16.51 \pm 1.56$ |
| (5) | LeMul (proposed) | $0.834 \pm 0.169$ | $\mathbf{15.49 \pm 1.50}$ |

BFM results comparison with baselines.

# THANK YOU

https://github.com/VinAIResearch/LeMul