SELF-SUPERVISED INVERSE RENDERING

WILL SMITH (+ YE YU & TATSURO KOIZUMI)

UNIVERSITY of York

INδAM

10TH FEBRUARY 2021

ROYAL ACADEMY OF ENGINEERING

LEVERHULME TRUST

OVERVIEW

- Learning inverse rendering without direct supervision

1. InverseRenderNet: Outdoor, scene level inverse rendering
   - Self-supervised by differentiable rendering

2. “Backwards rasterisation”: faces, using a 3D morphable model
   - Towards avoiding forward rendering

SCENE LEVEL, OUTDOOR INVERSE RENDERING

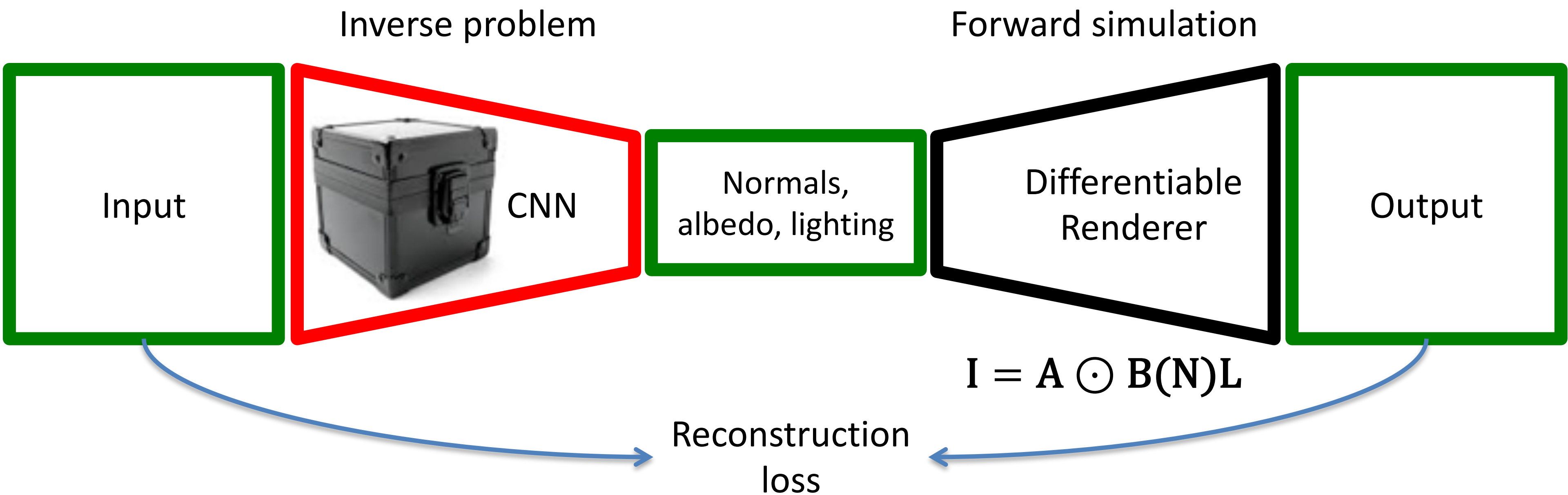
- Geometry
  - Surface normal?
  - Depth map?
  - Mesh?
  - Implicit surface?
- Material properties
  - Diffuse albedo?
  - Specular params?
- Illumination
- Shadows

BEYOND SUPERVISION

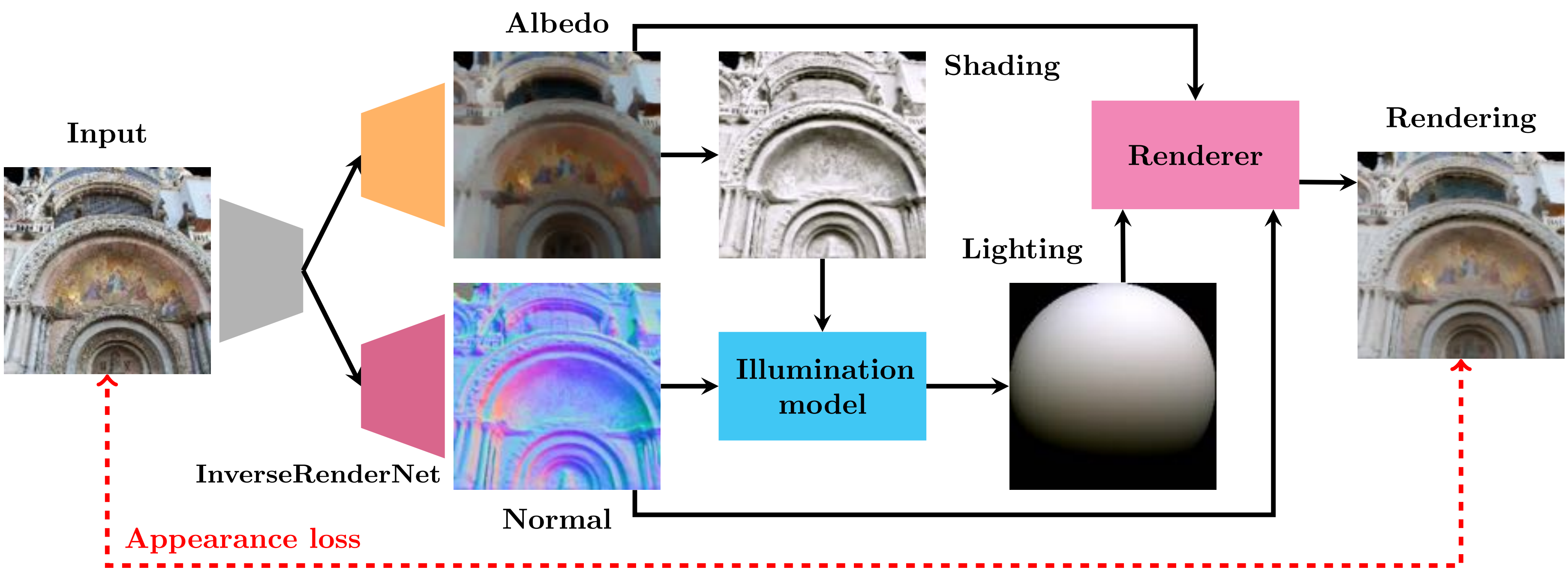
- What if it's very difficult (or impossible) to obtain training data and/or annotations?
- What if the inverse problem we're trying to learn is unsolved?

1. Use output of an existing algorithm
   - But then just learning to replicate performance
2. Synthesise images with known ground truth
   - Generalisation limited by diversity/realism of data

SELF-SUPERVISION



INVERSERENDERNET



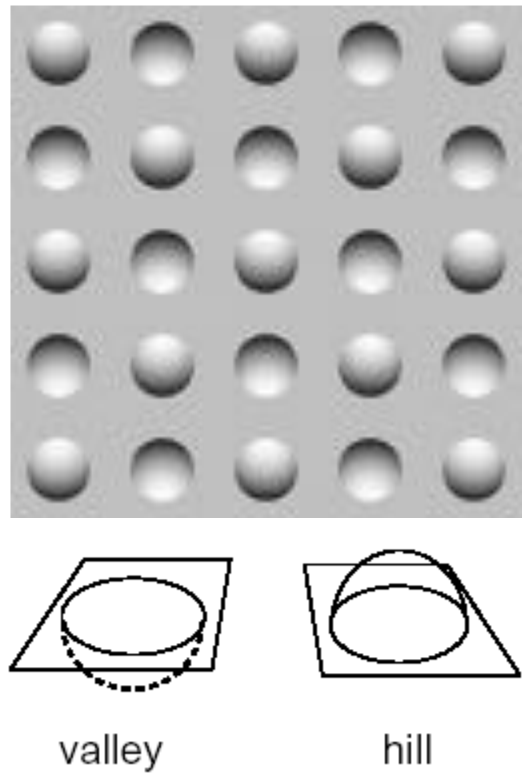
Y. Yu and **W.A.P. Smith**. InverseRenderNet: Learning single image inverse rendering. In Proc. CVPR, 2019.

ILL-POSED PROBLEM: SHADED VERSUS PAINTED

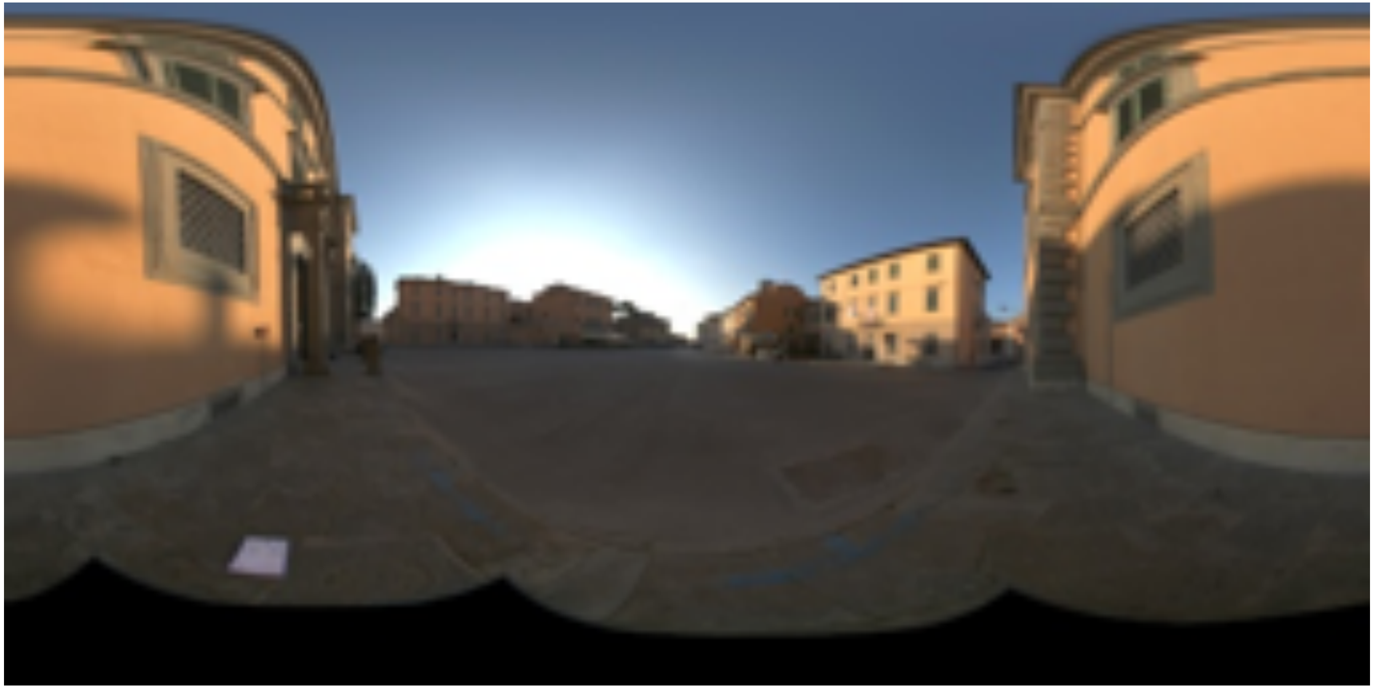
ILL-POSED PROBLEM: SHADED VERSUS PAINTED

We need more supervision!

SHAPE-FROM-SHADING IN HUMANS



NATURAL ILLUMINATION

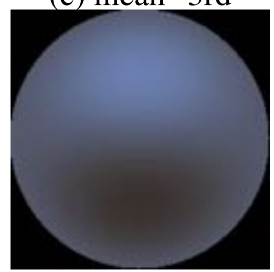


STATISTICAL ILLUMINATION MODEL

(a) mean + 1st

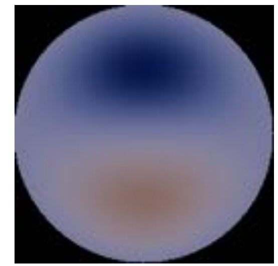
(b) mean + 2nd

(c) mean - 3rd



(d) mean

(e) mean + 3rd

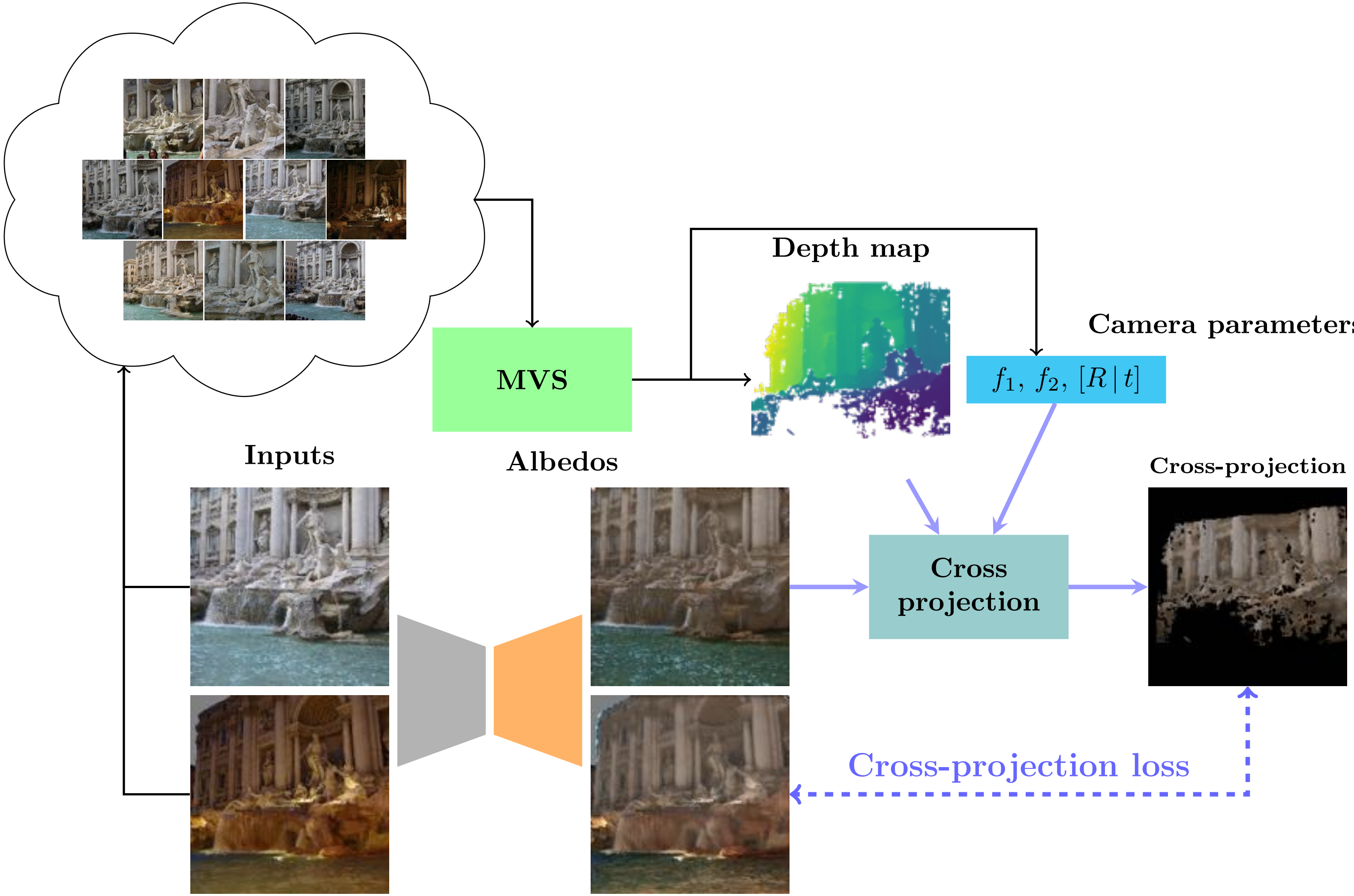


(f) mean - 2nd

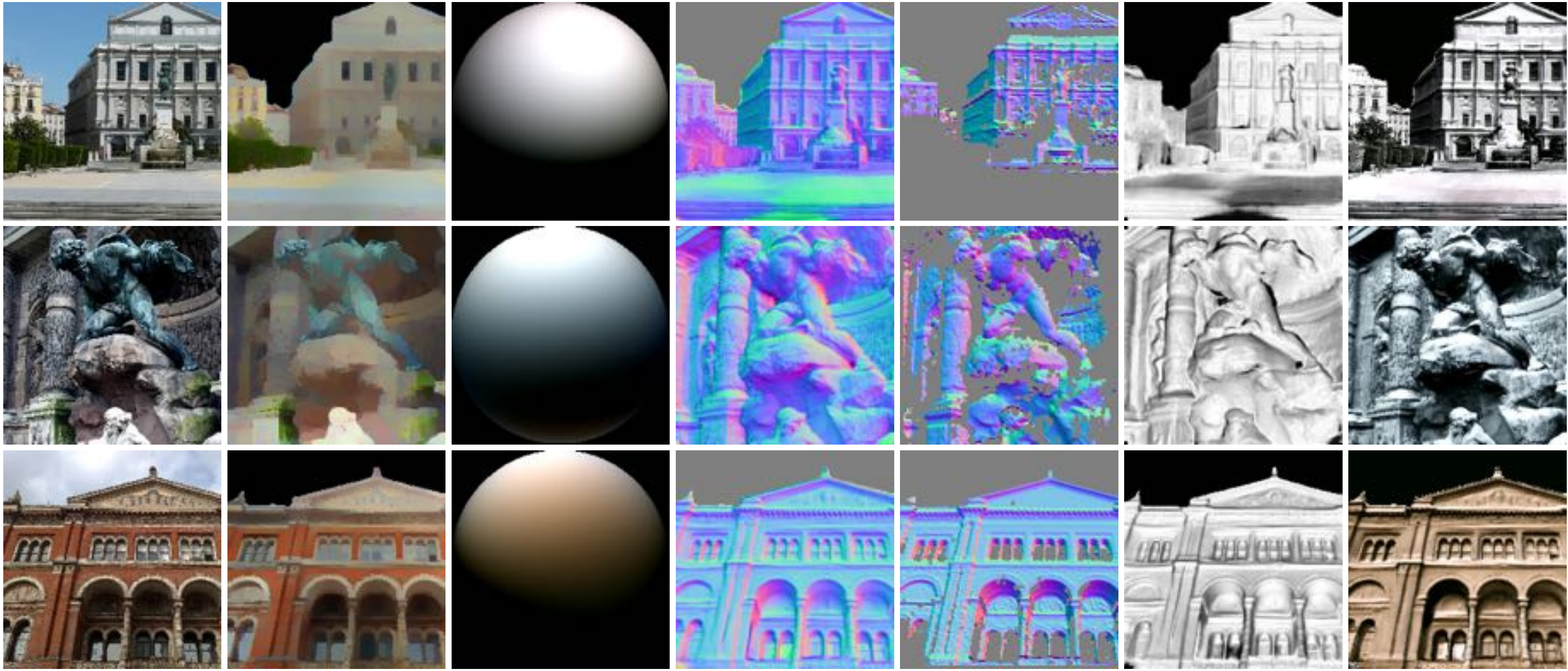
(g) mean - 1st

MULTIVIEW SUPERVISION

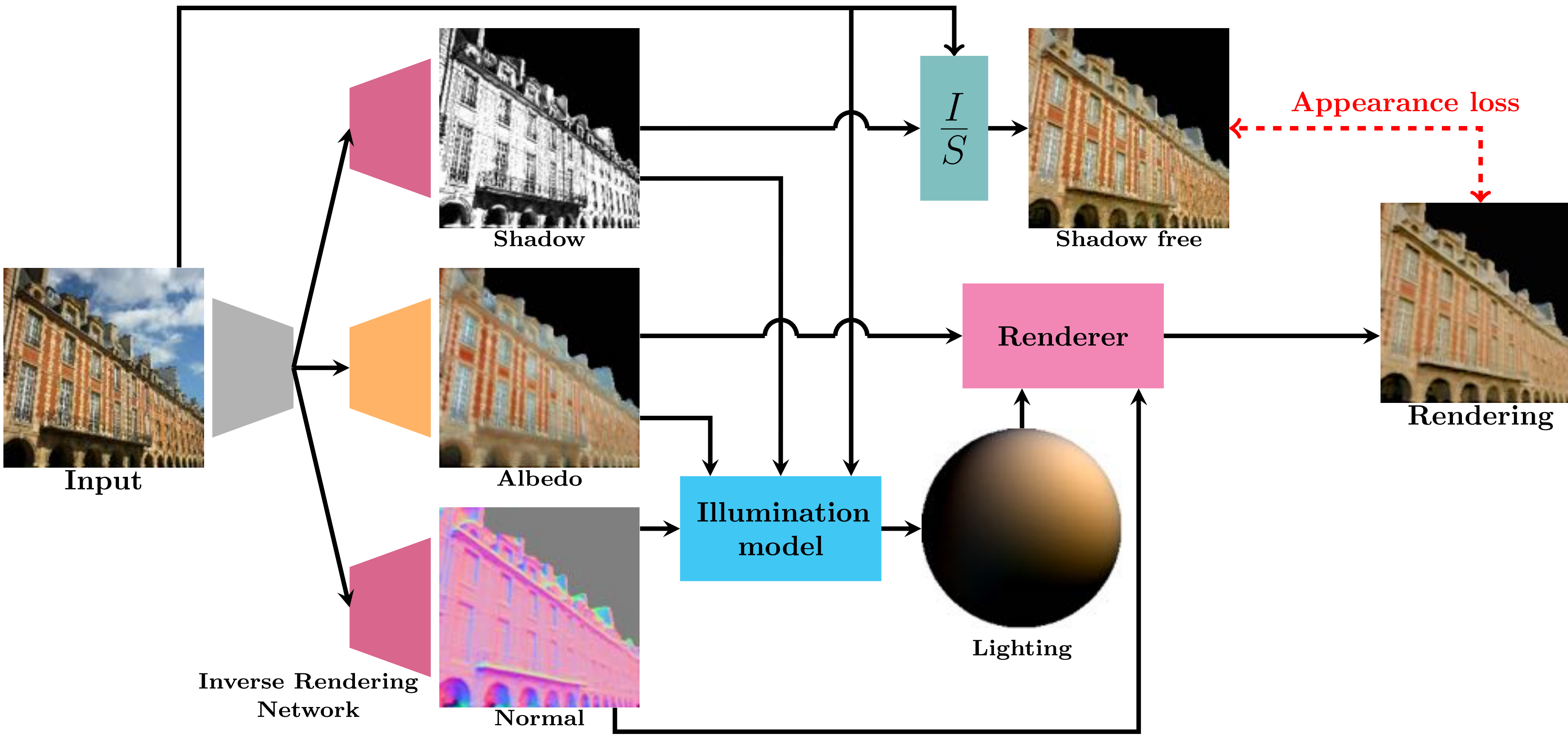
MULTIVIEW SUPERVISION



INVERSE RENDERING RESULTS

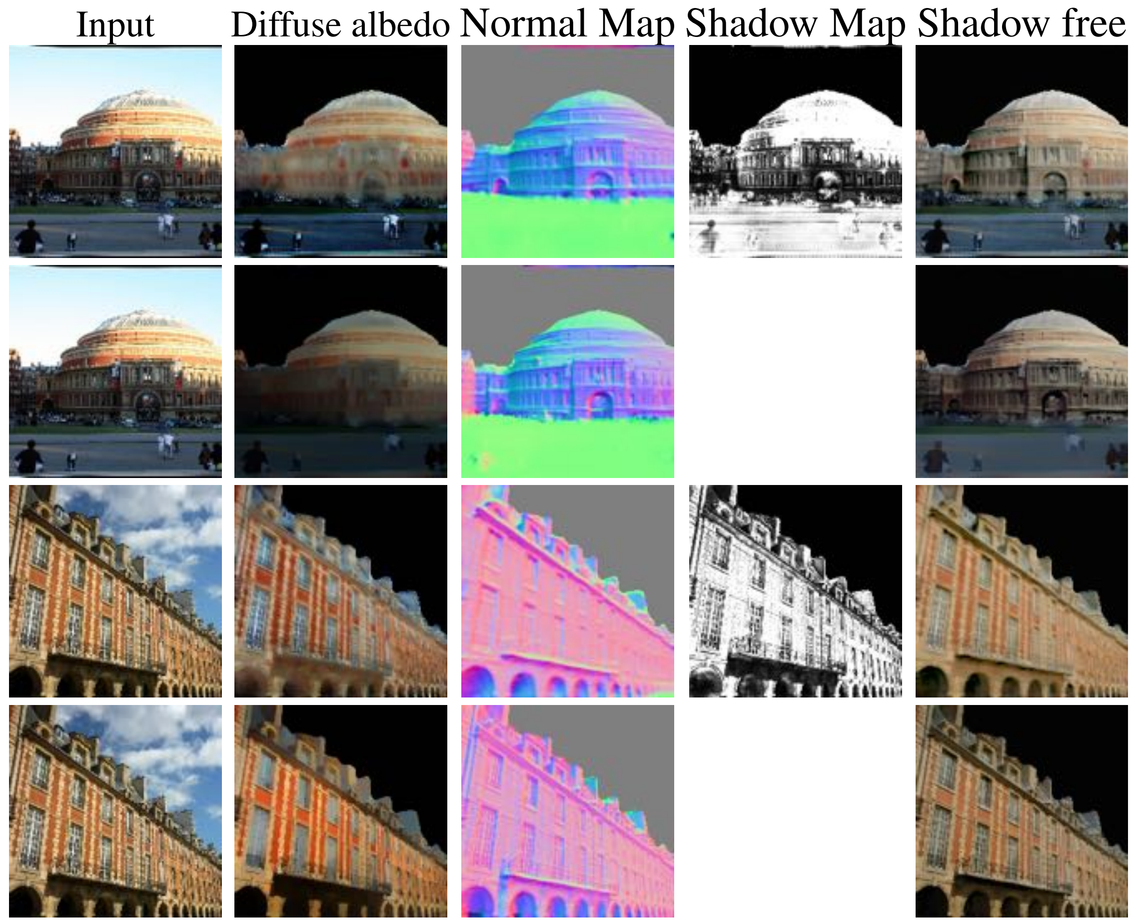


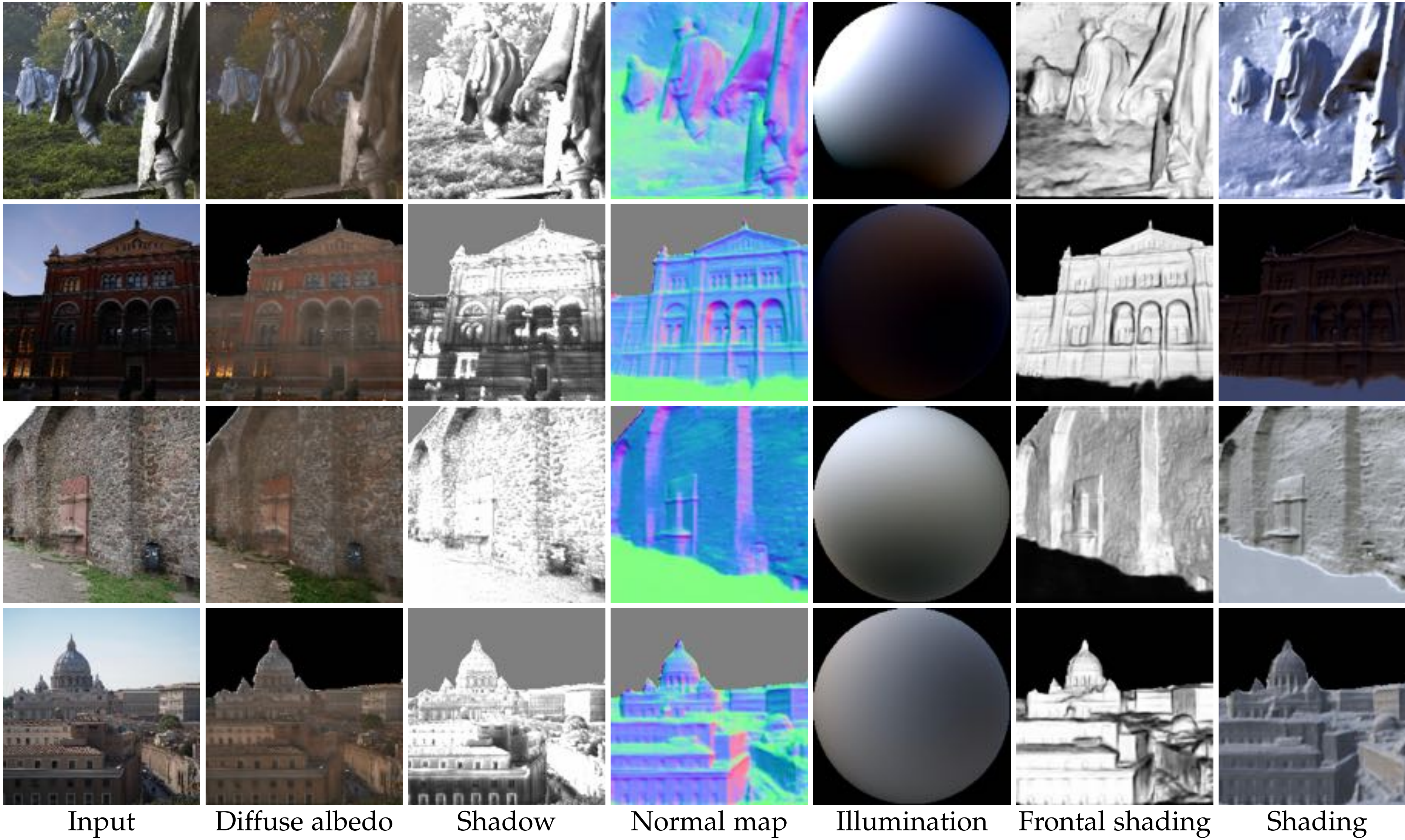
INVERSERENDERNET++

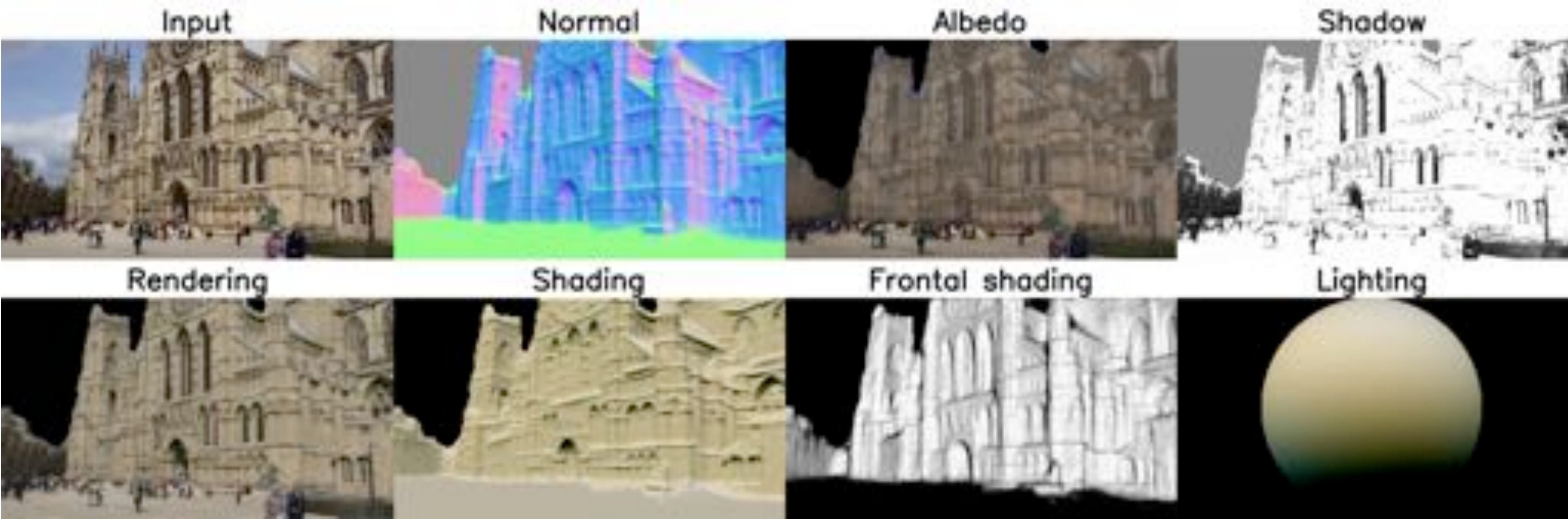


Y. Yu and **W.A.P. Smith**, Outdoor inverse rendering from a single image using multiview self-supervision, *IEEE T-PAMI*, to appear.

SHADOW ESTIMATION

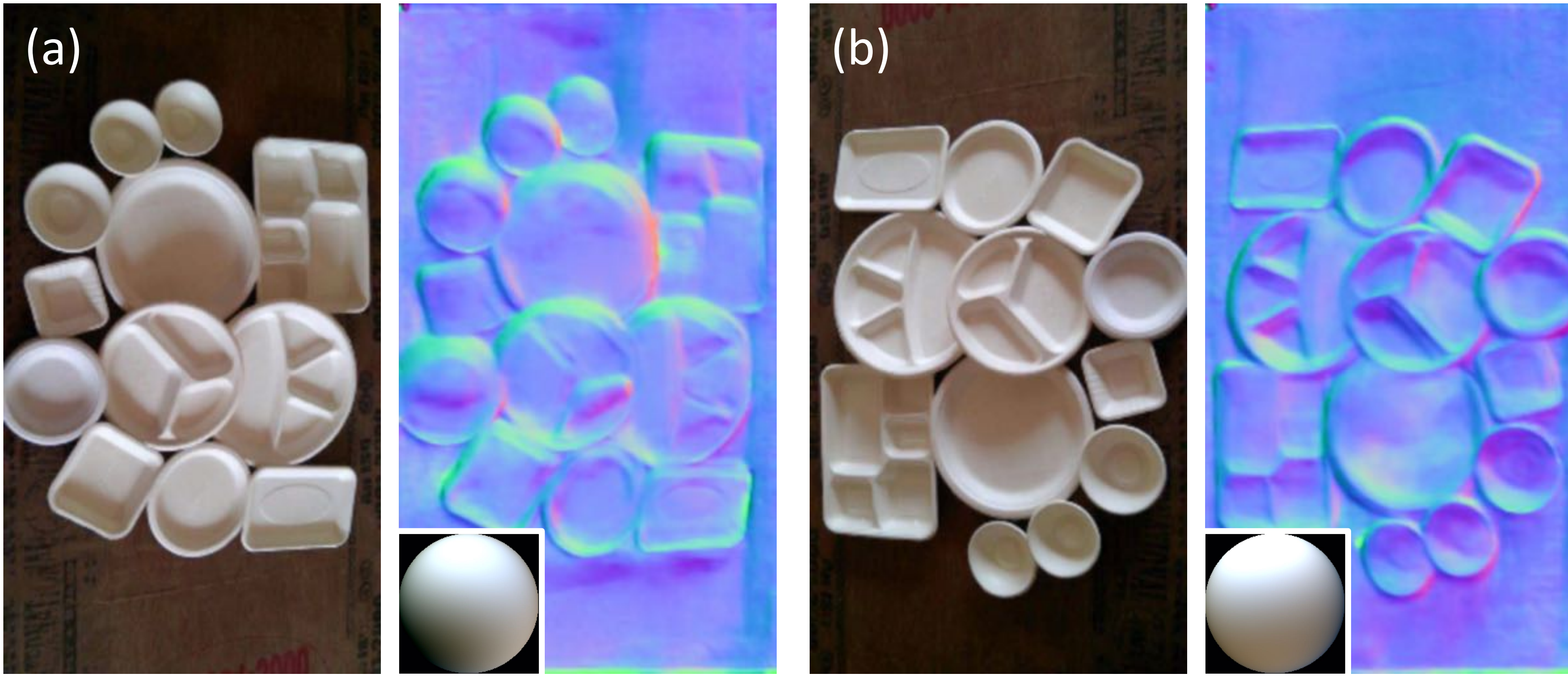




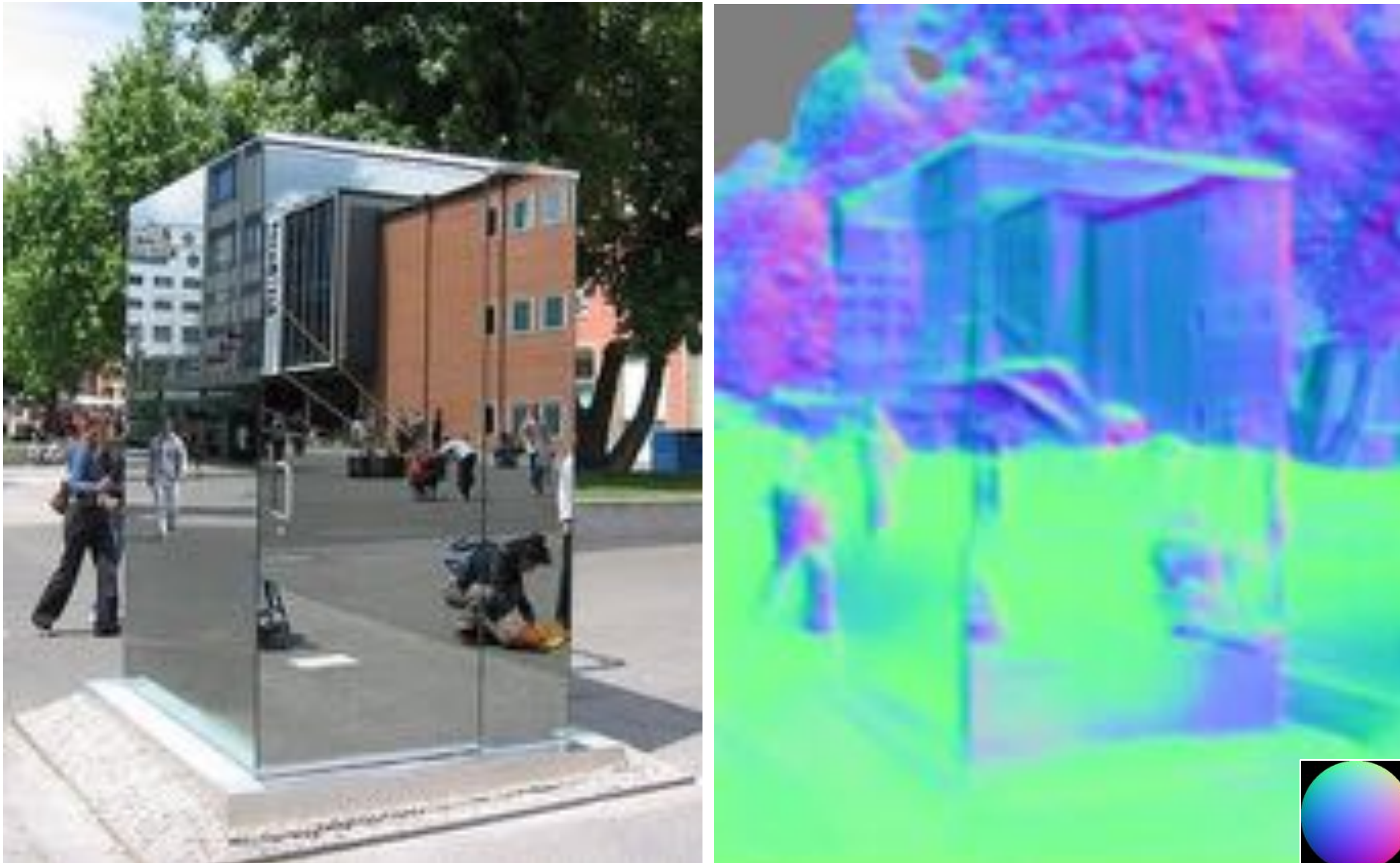

Input
Normal
Albedo
Shadow
Rendering
Shading
Frontal shading
Lighting

WHAT DOES IT ACTUALLY LEARN?

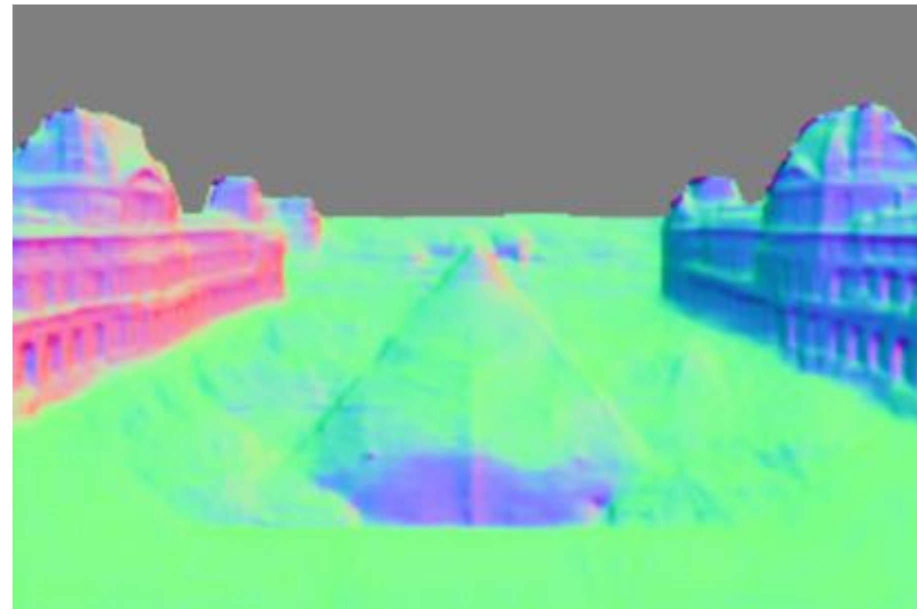
- Shape-from-...?
  - Shading?
  - Texture?
  - Shadows?
  - Ambient occlusion?
  - ...
  - Semantics?
- General principles of shape-from-X?

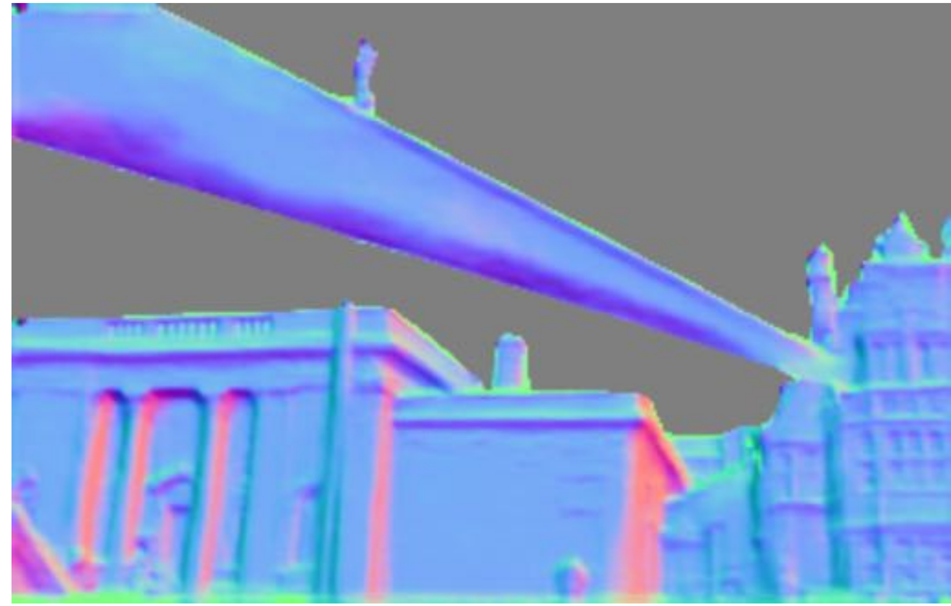


(a)
(b)



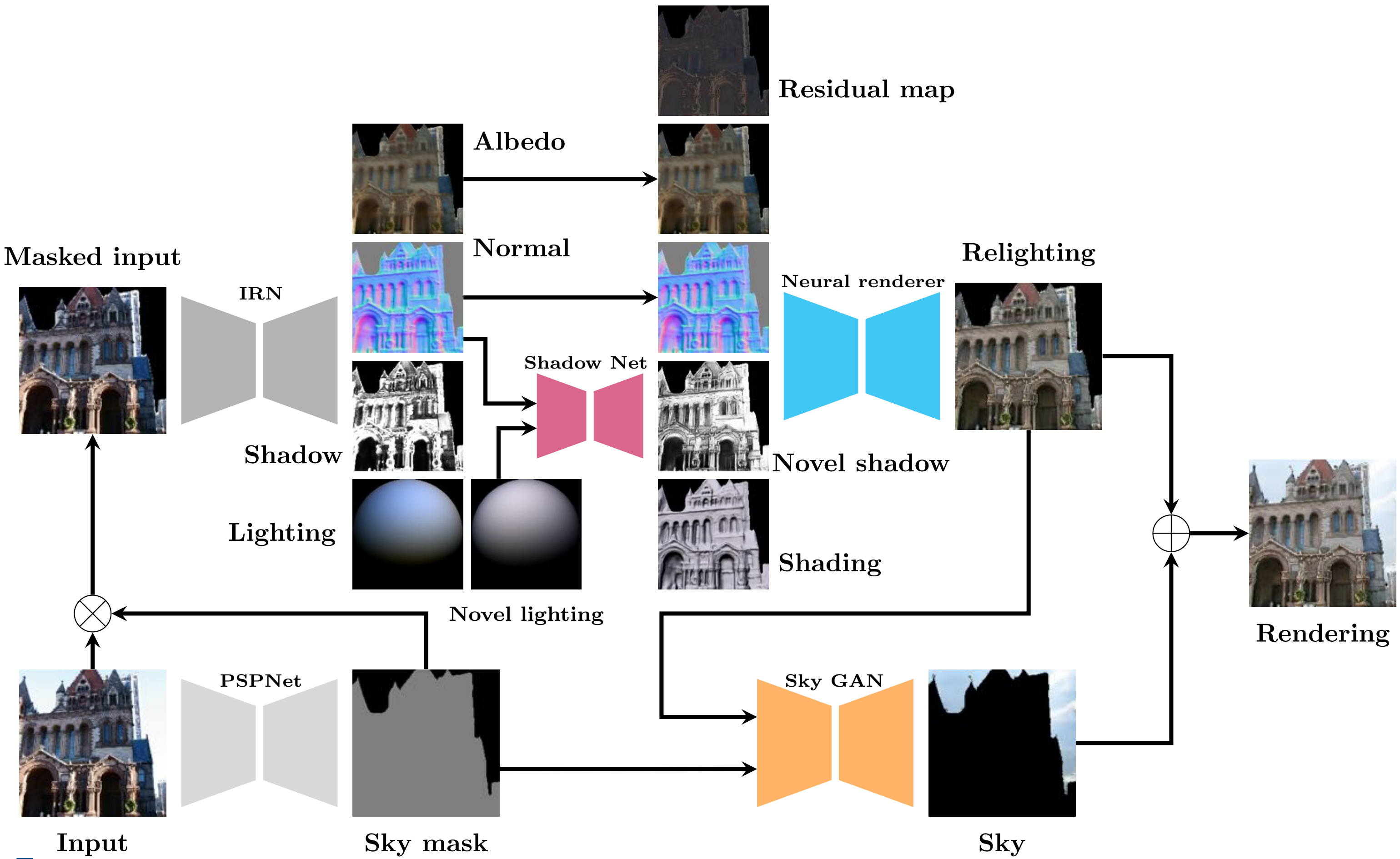

AU FIL DE LOIRE

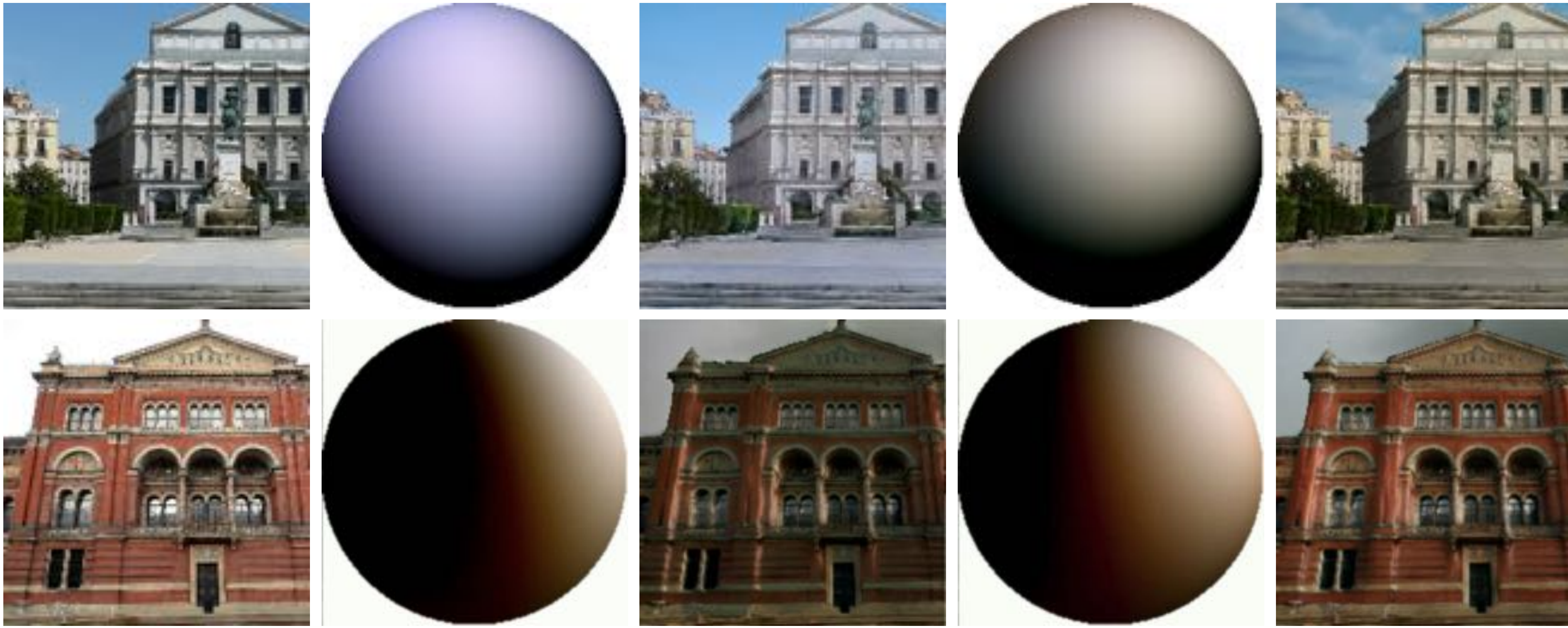
APPLICATION: RELIGHTING

RELIGHTING WITH NEURAL RENDERING



Y. Yu, A. Meka, M. Elgharib, H.-P. Seidel, C. Theobalt, **W.A.P. Smith**, Self-supervised Outdoor Scene Relighting, In Proc. ECCV, 2020.

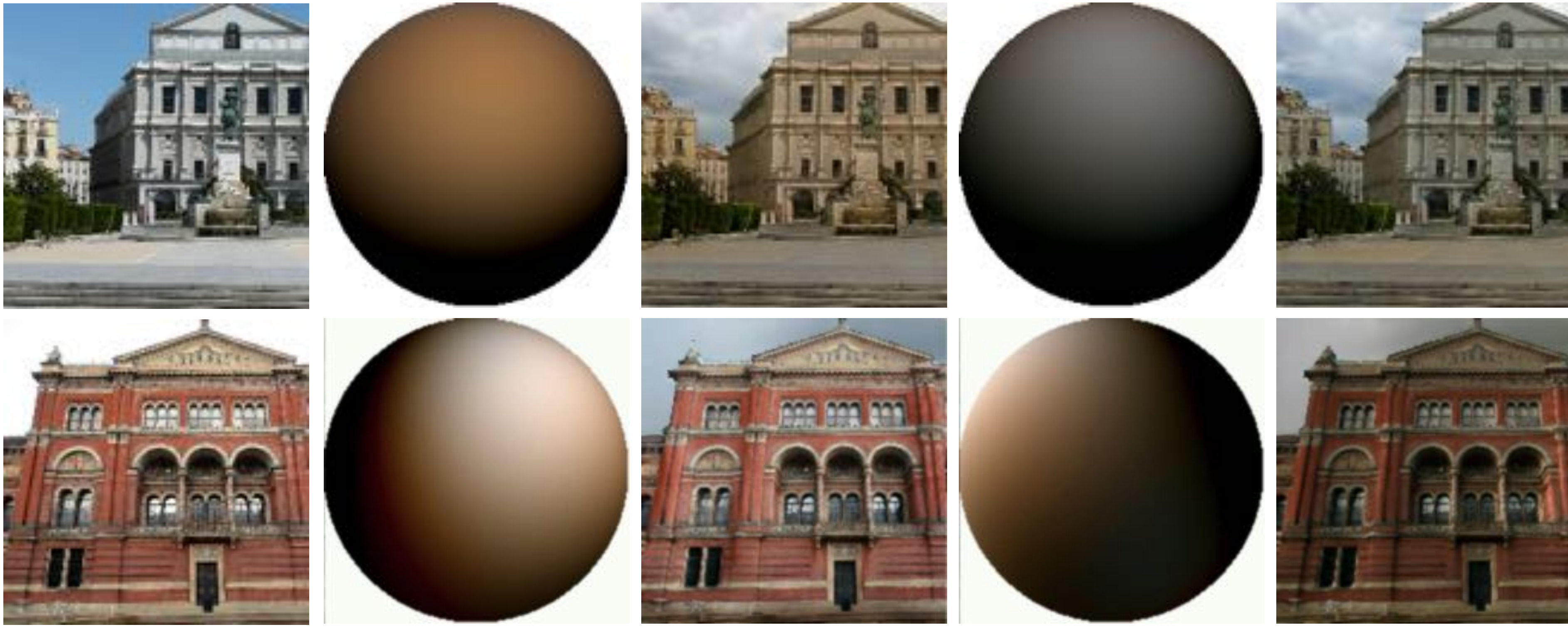
RELIGHTING WITH NEURAL RENDERING



Input Illumination1 Relighting1 Illumination2 Relighting2

iNδAM

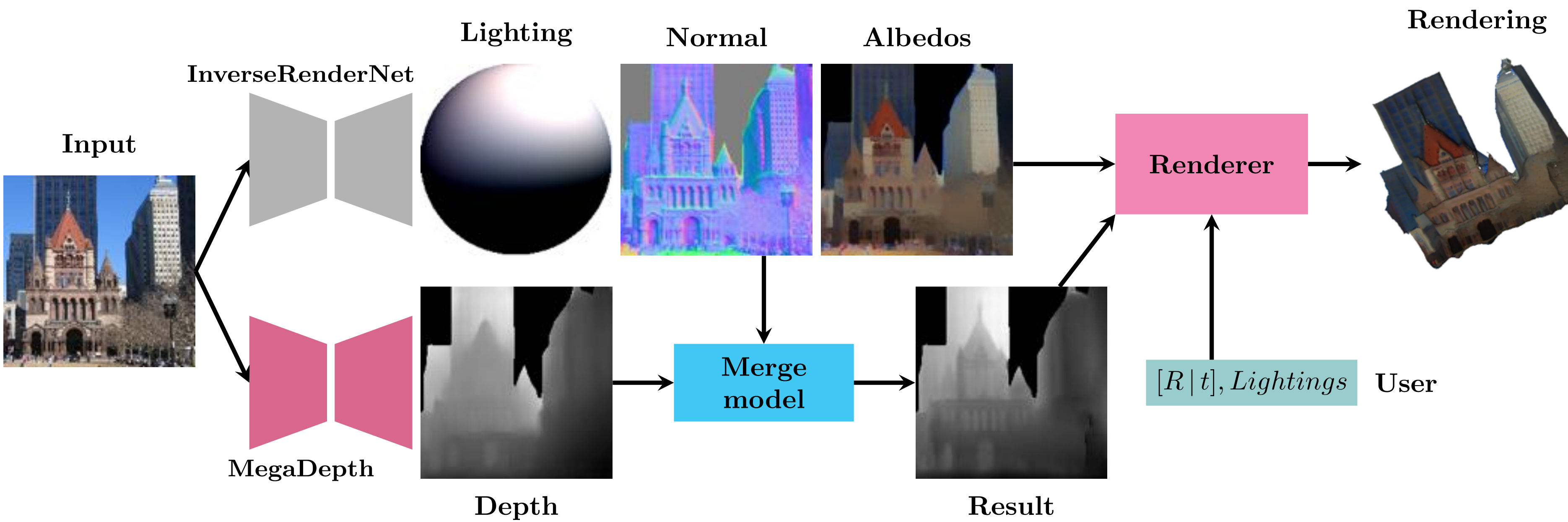
RELIGHTING WITH NEURAL RENDERING



Input Illumination3 Relighting3 Illumination4 Relighting4

RELIGHTING WITH NEURAL RENDERING

MERGING WITH MONODEPTH

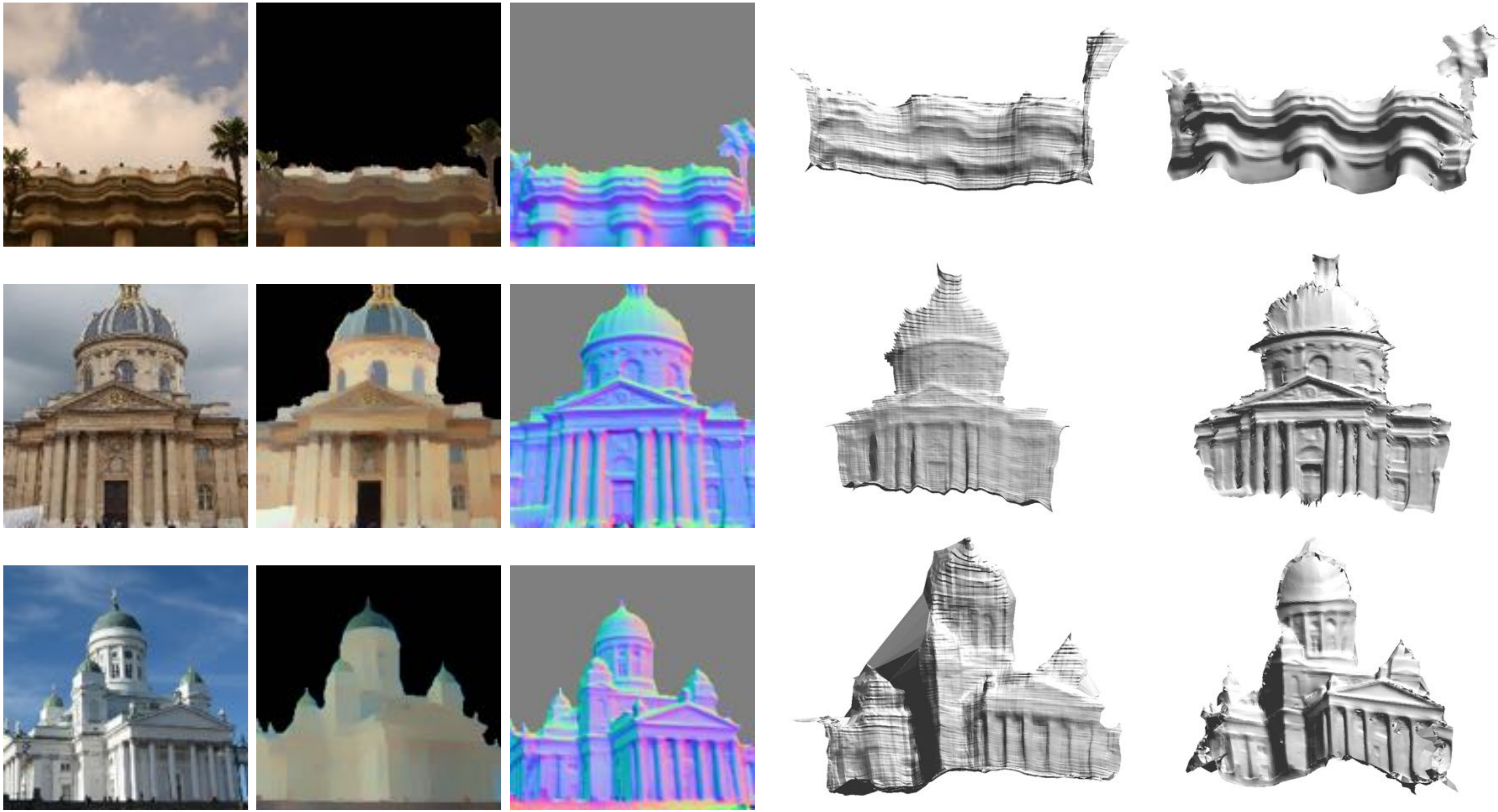


Y. Yu and **W.A.P. Smith**. Depth estimation meets inverse rendering for single image novel view synthesis. In Proc. CVMP, 2019.

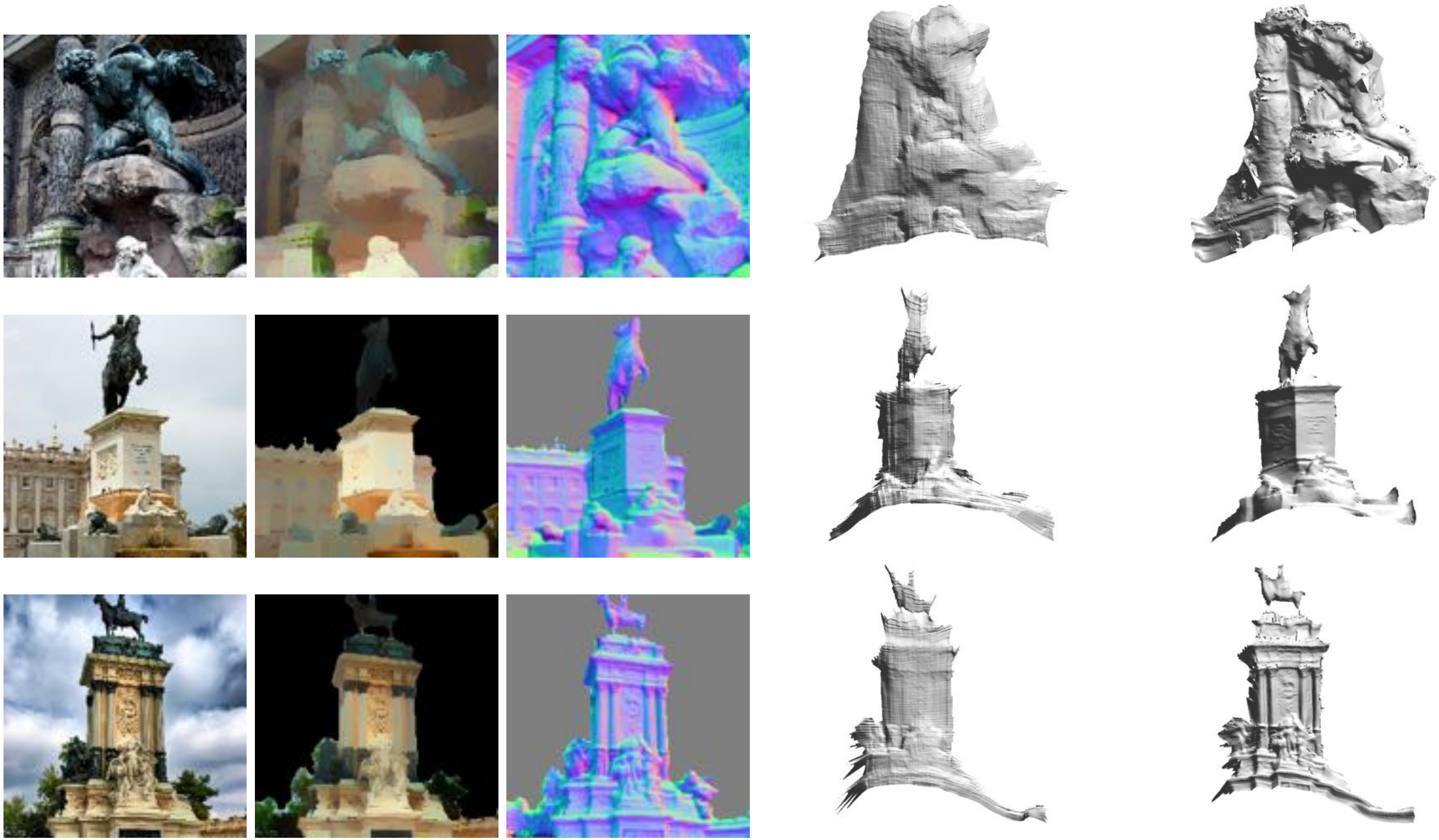
Z. Li and N. Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In Proc. CVPR, 2018.

D. Nehab, S. Rusinkiewicz, J. Davis and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. ACM TOG, 2005.

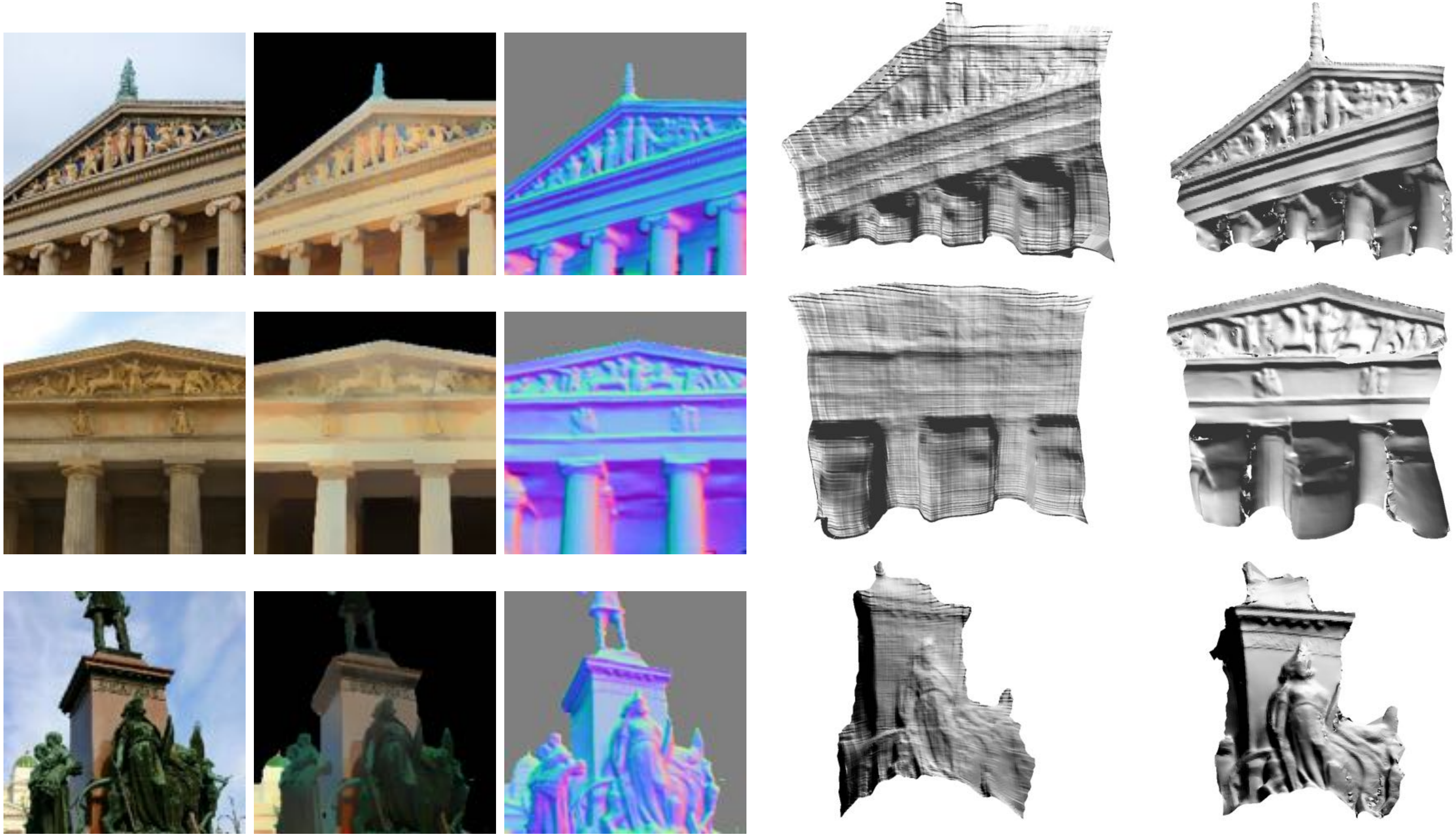
RESULTS



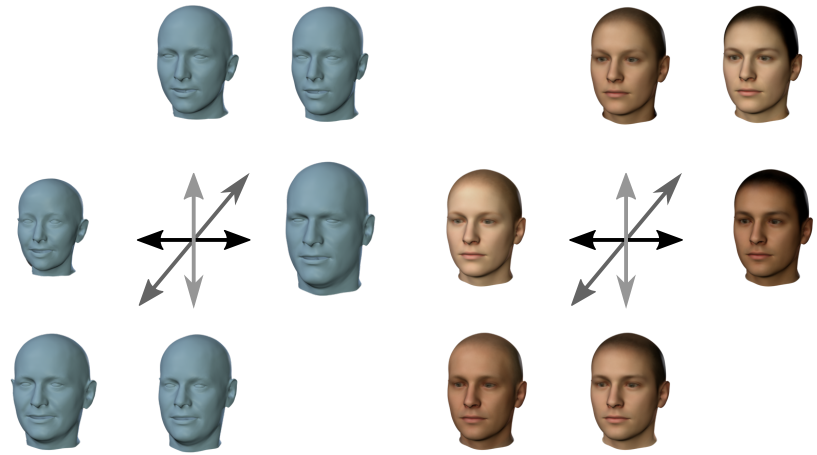
RESULTS

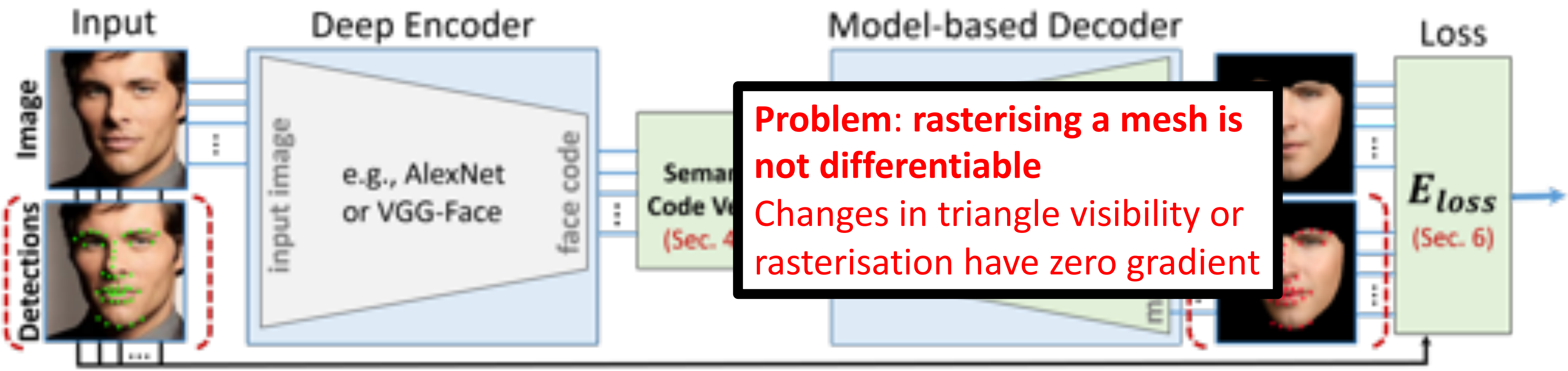


RESULTS



CLASS SPECIFIC METHODS



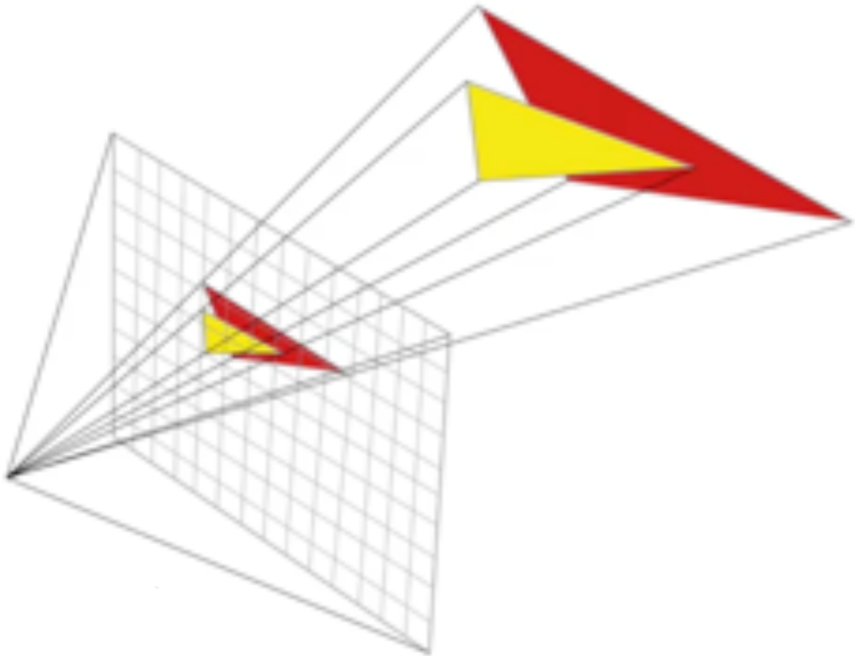


Problem: rasterising a mesh is not differentiable

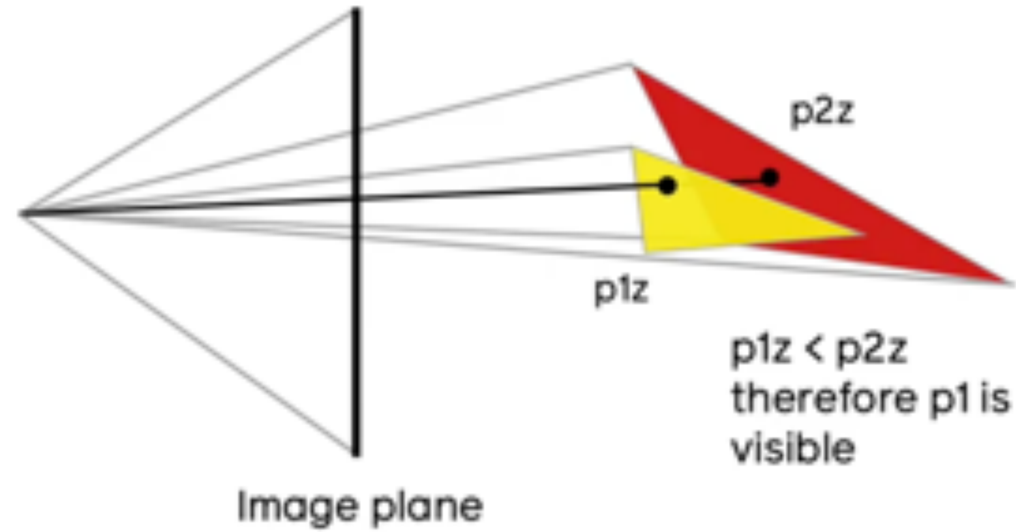
Changes in triangle visibility or rasterisation have zero gradient

A. Tewari et al. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In Proc. ICCV, 2017.

DISCONTINUITIES IN RENDERING



Rasterisation



Visibility

BACKWARDS RASTERISATION

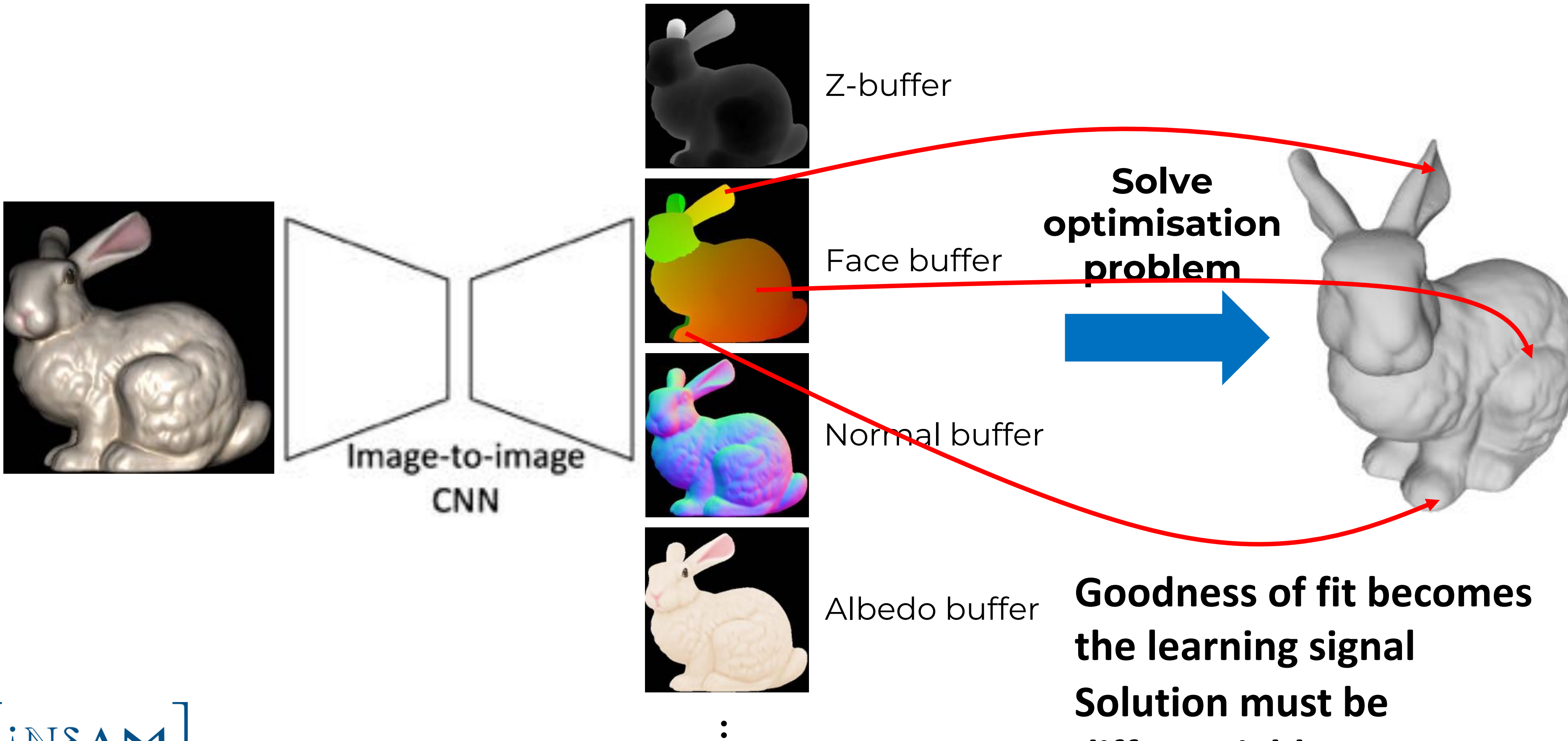
Rasterisation = Given a mesh…

- For every pixel, find closest mesh triangle that covers the pixel
- Having established correspondence from mesh model to image, compute a colour from other rasterised quantities (depth, normal, albedo etc)

Backwards Rasterisation = Given an image…

- Predict the buffers that would have arisen from rasterising the model
- Solve optimisation to find model consistent with predicted buffers

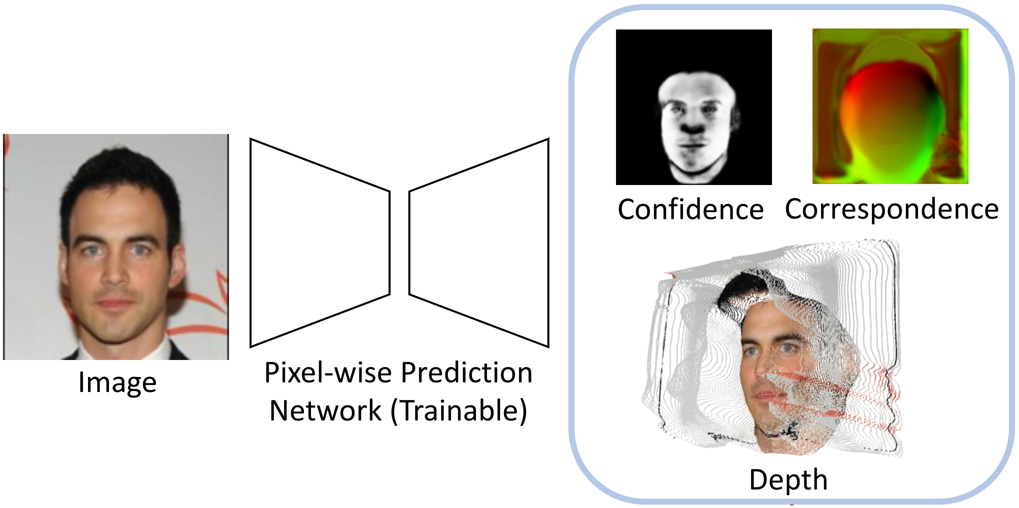
BACKWARDS RASTERISATION



Goodness of fit becomes the learning signal

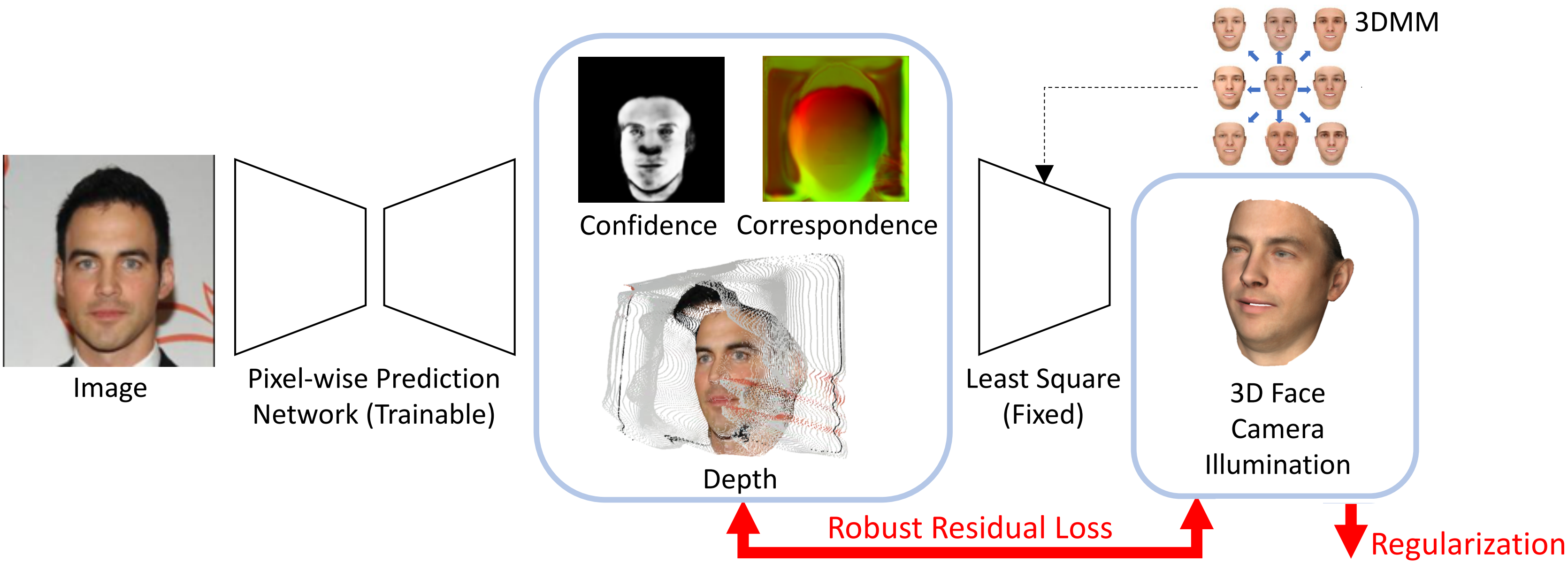
Solution must be differentiable

BACKWARDS RASTERISATION



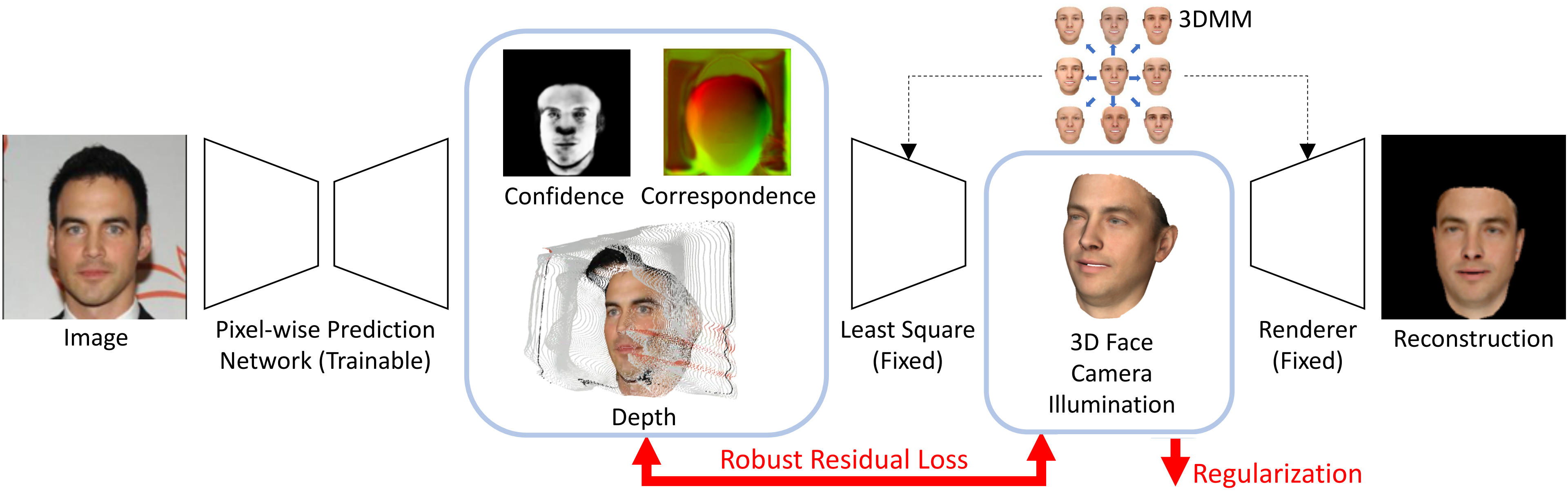
T. Koizumi and **W.A.P. Smith**, "Look Ma, no landmarks!" - Unsupervised, model-based dense face alignment, *Proc. ECCV*, 2020.

BACKWARDS RASTERISATION



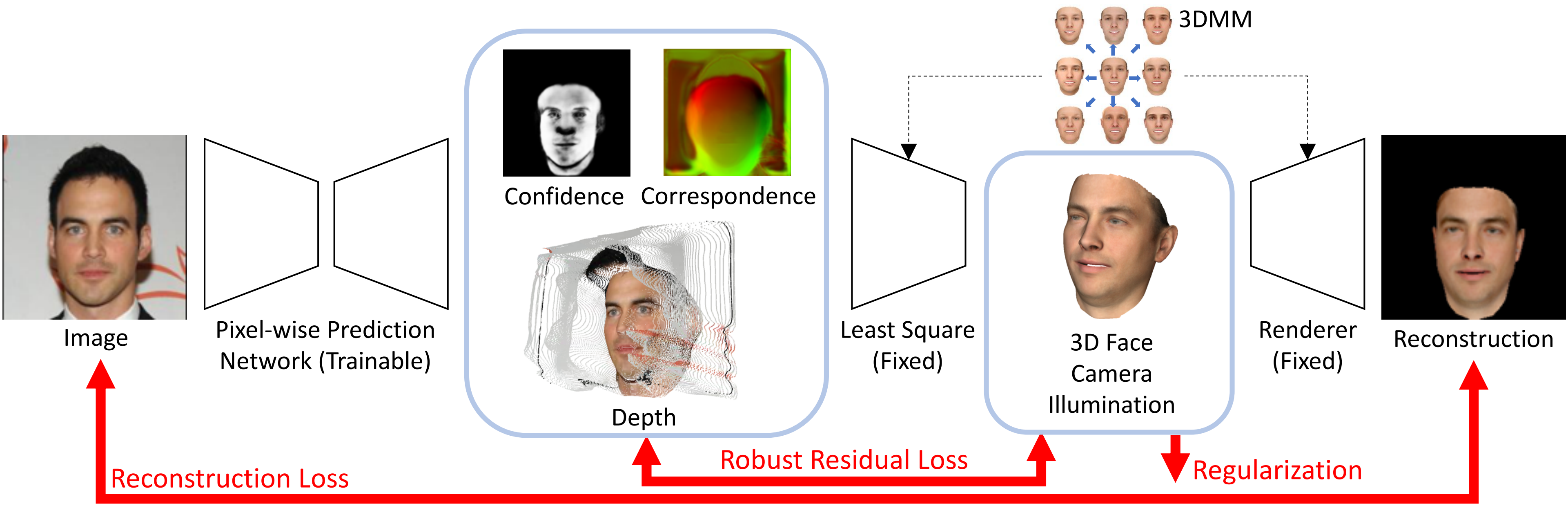
T. Koizumi and **W.A.P. Smith**, "Look Ma, no landmarks!" - Unsupervised, model-based dense face alignment, *Proc. ECCV*, 2020.

BACKWARDS RASTERISATION



T. Koizumi and **W.A.P. Smith**, "Look Ma, no landmarks!" - Unsupervised, model-based dense face alignment, *Proc. ECCV*, 2020.

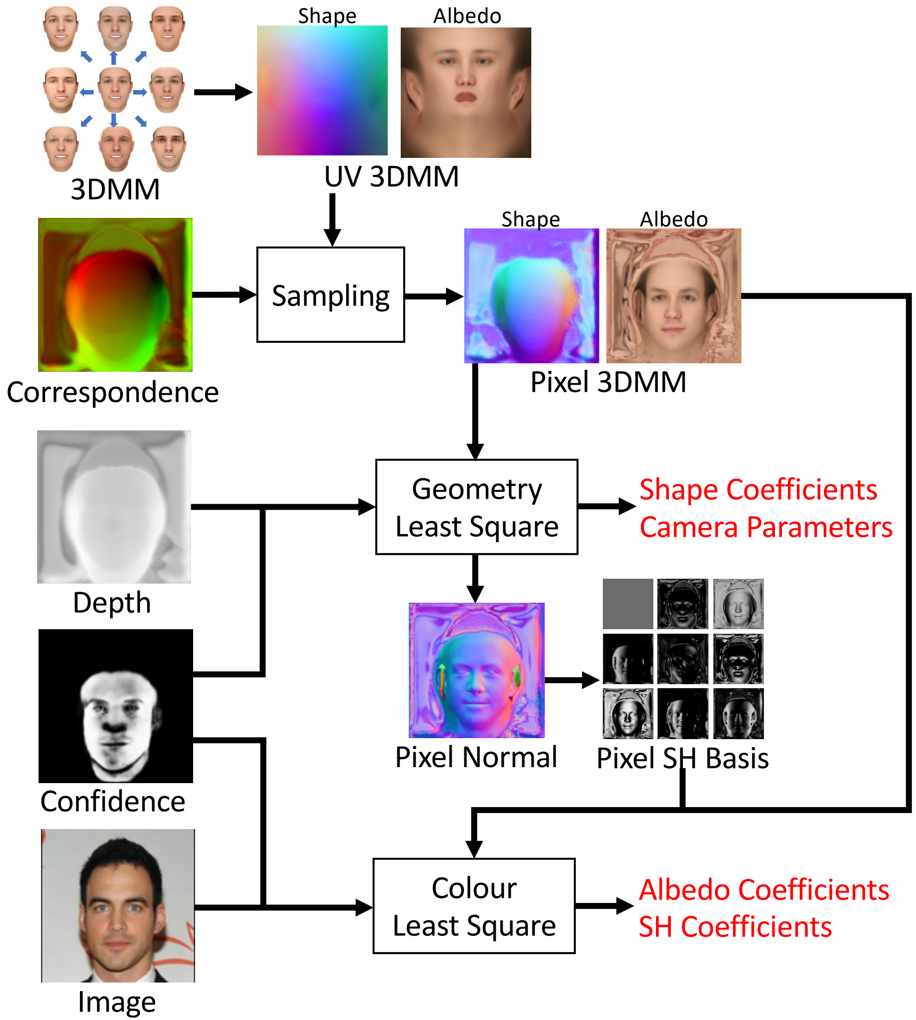
BACKWARDS RASTERISATION



T. Koizumi and **W.A.P. Smith**, "Look Ma, no landmarks!" - Unsupervised, model-based dense face alignment, *Proc. ECCV*, 2020.

iNδAM

LINEAR LEAST SQUARES FITTING



RATIONALE

1. Minimal representation
   - Compute geometric parameters from correspondence
   - Compute photometric parameters from image + geometric parameters
2. Task better suited to CNN architecture, smaller network
3. Every pixel can contribute to appearance losses – alternative to soft rasterization
4. Defer estimation of actual face geometry – intermediate representation
5. Can train completely unsupervised – no landmarks!

INVERSE SPHERICAL HARMONIC LIGHTING

$$\mathbf{i} = \mathbf{r} \odot \overbrace{\mathbf{B}(\mathbf{n})\mathbf{L}}^{\text{Shading}}$$

Appearance

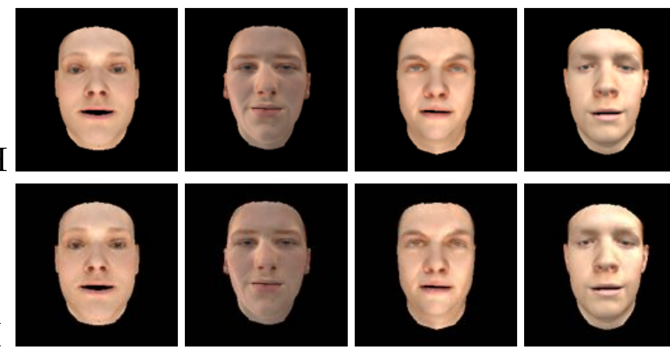
Reflectance (albedo)

SH basis (depends on normal)

Lighting coefficients

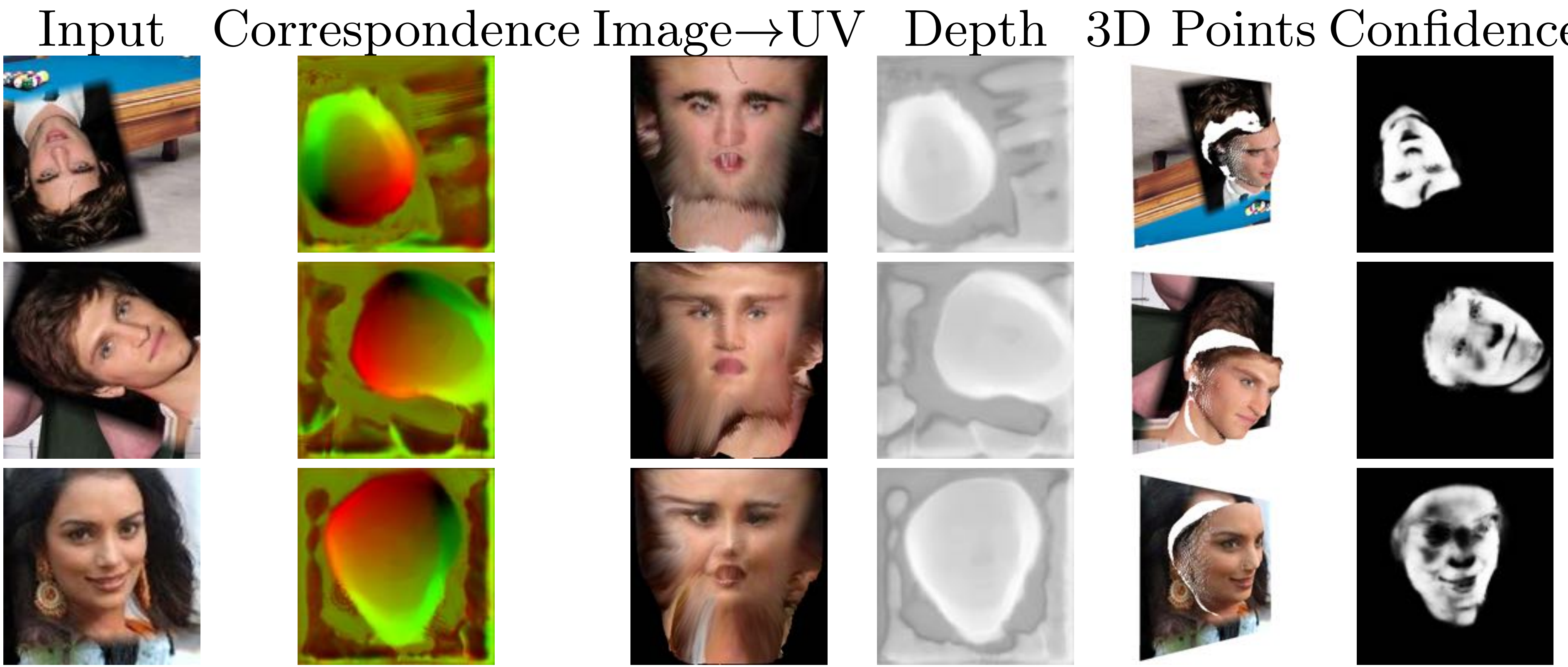
$$\overbrace{\mathbf{i} \oslash \mathbf{B}(\mathbf{n})\mathbf{L}}^{\text{Inverse shading}} = \mathbf{r}$$

Closed form (linear least squares) solution for albedo and lighting parameters simultaneously!

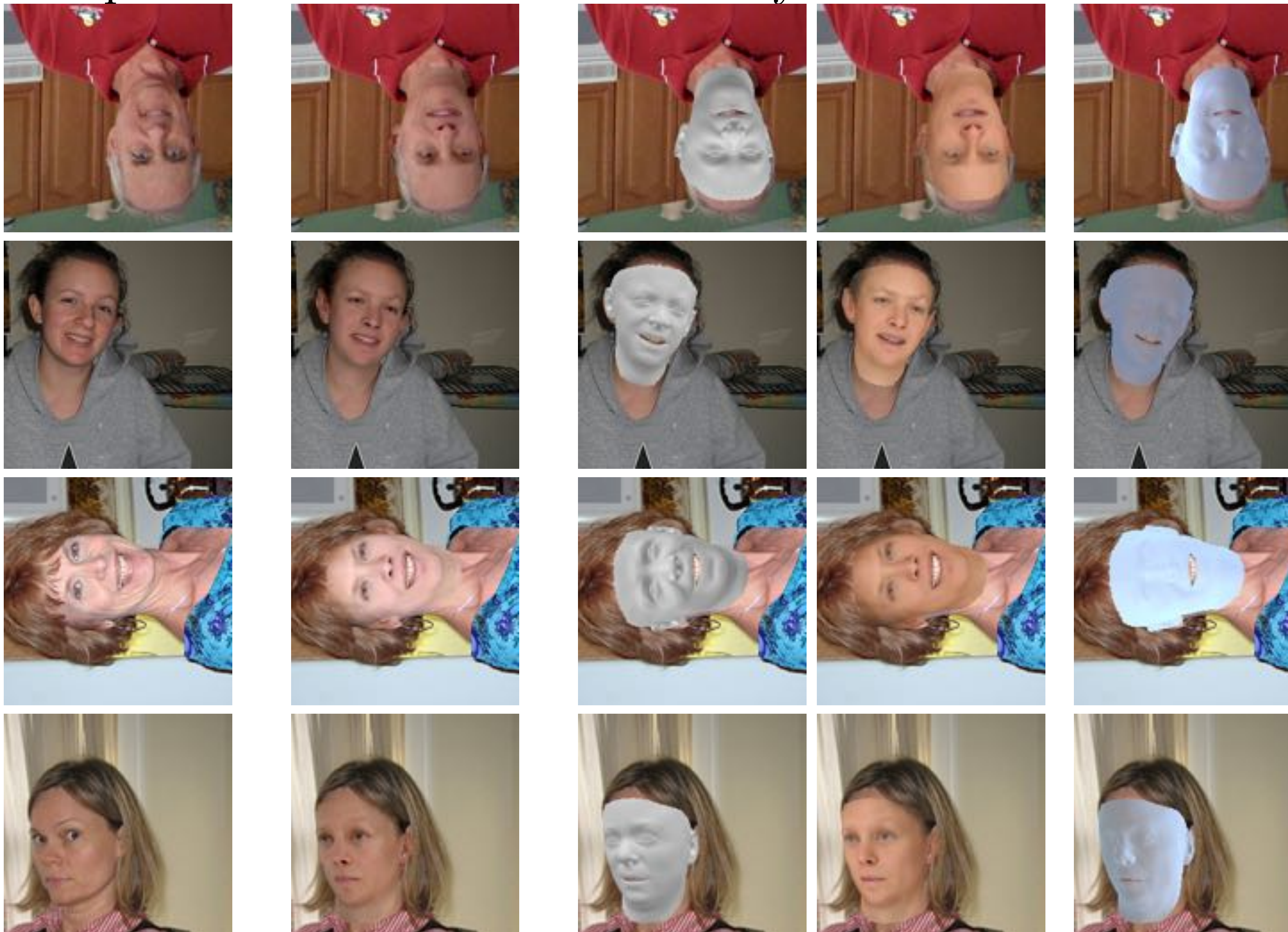


Regular SH Inverse SH				
Max error	0.049	0.109	0.123	0.272
RMS error	0.019	0.043	0.029	0.058

RESULTS



Input | Reconstruction | Geometry | Albedo | Illumination



MULTIFRAME AGGREGATION

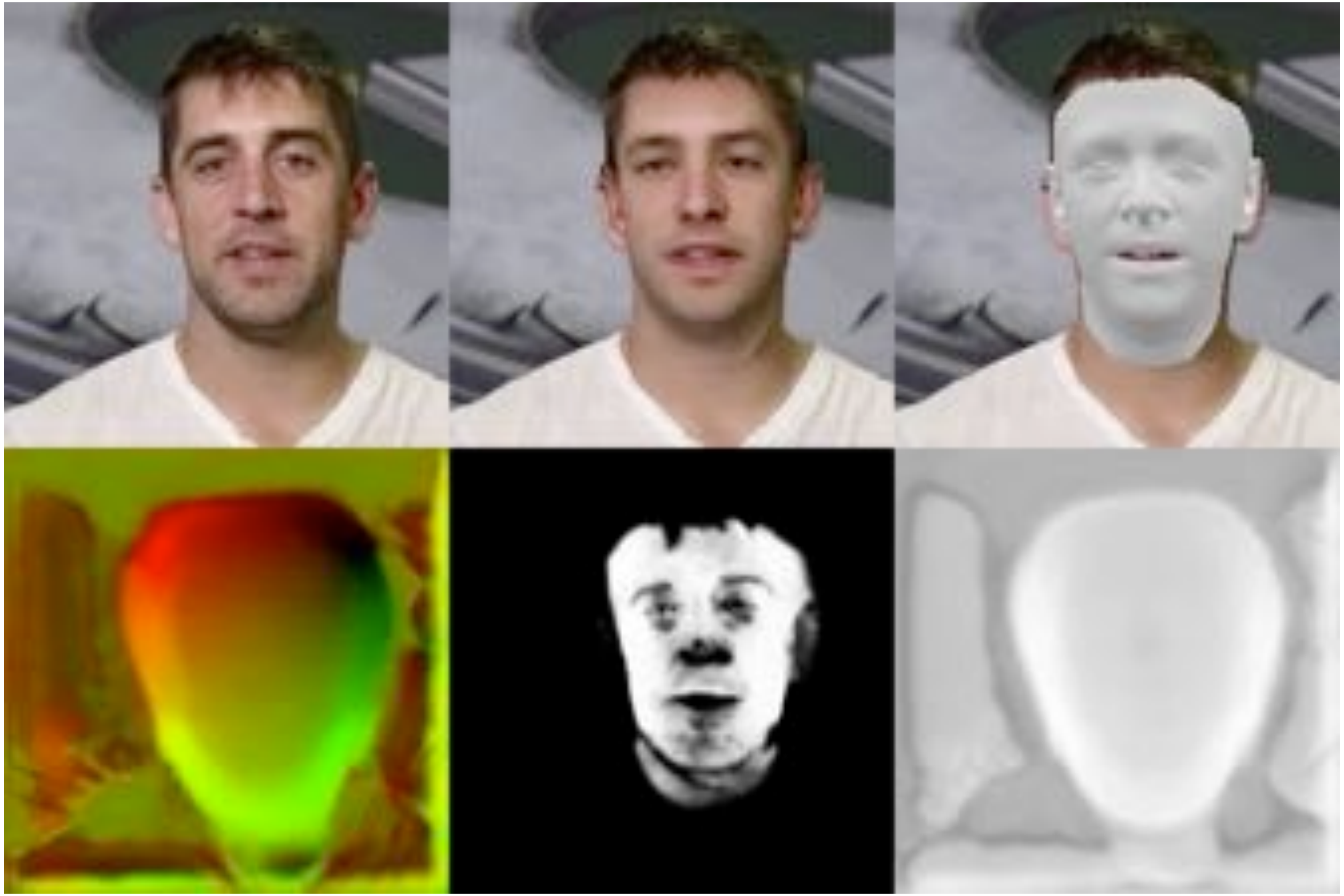
Input

Single frame

Multi frame

iNδAM

VIDEO FITTING RESULTS



CONCLUSIONS

- "Models" (physics-based reflectance models, statistical object class models, geometric models from MVS, linear least squares fitting) can supervise learning
- The network "learns" from the model
- The model encapsulates what we know about the world
- All models are wrong
  - Should reflectance/rendering models be partially (fully?) learnable?
  - Broader question: what is the right balance between "modelling" (human understanding/domain knowledge) and learning