Image Factorization and Manipulation with Generative Regularizations

Zhixiang Wang
PhD candidate
The University of Tokyo

Goal: GenAI + Advanced Cameras for VFX

Reduce actor, time, and money costs

Research Works

Special Hardwares

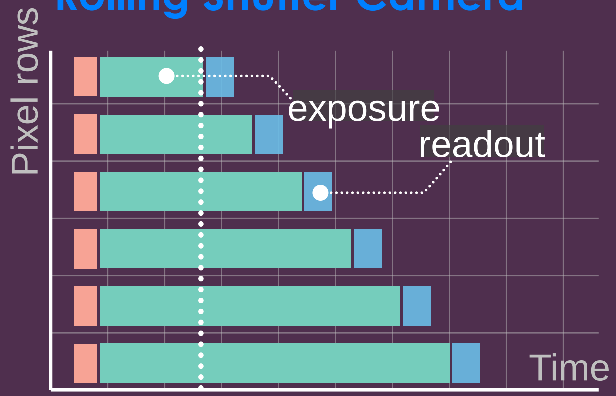
Polarimetric Camera

Wang et al, CVPR 2019

Infrared Camera

Wang et al, CVPR 2019
Wei, **Wang** et al, AAAI'23

Rolling Shutter Camera



Wang et al, CVPR 2022
Ji, **Wang** et al, ICCV 2023

Foggy Scene Understanding

Ma, **Wang** et al, CVPR 2022

Generative Models

Geometric Distortion Correction

Wang et al, IJCV 2024

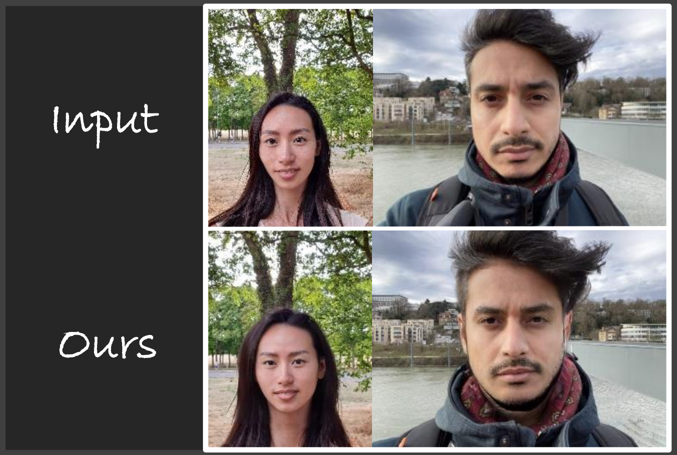
Background Replacement

Wang et al, SIGGRAPH 2024

Style Transfer

Chang, **Wang** et al, ECCV 2020

Viewpoint + Lens



Perspective Distortion Correction

Wang et al, IJCV 2024

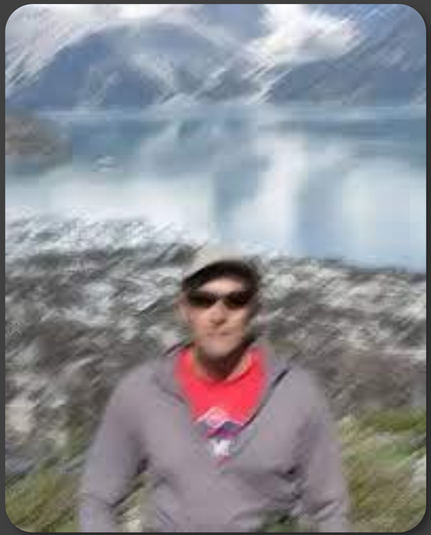
Background

Matting by Generation

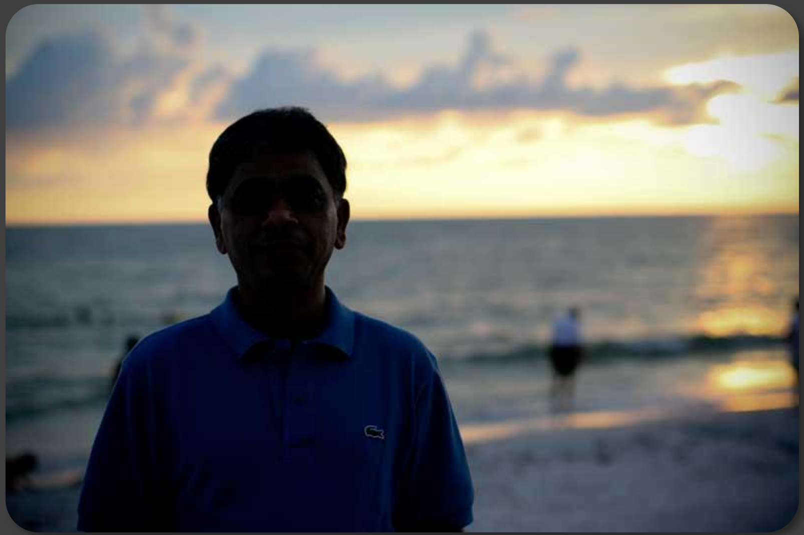
Wang et al, SIGGRAPH 2024

Good Photos are Not Easy to Take

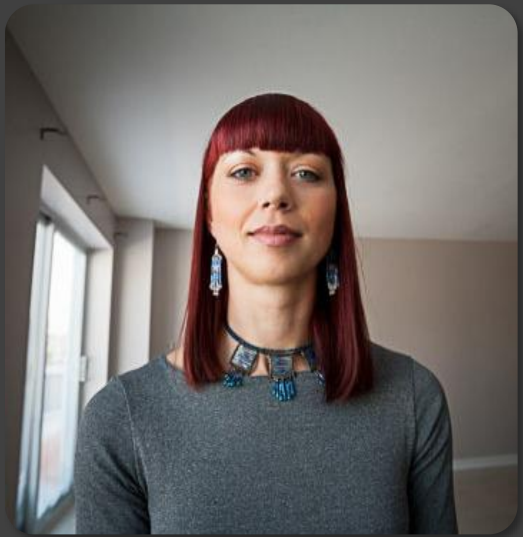
Examples of “bad/undesired” photos,
caused by unwanted imaging factors



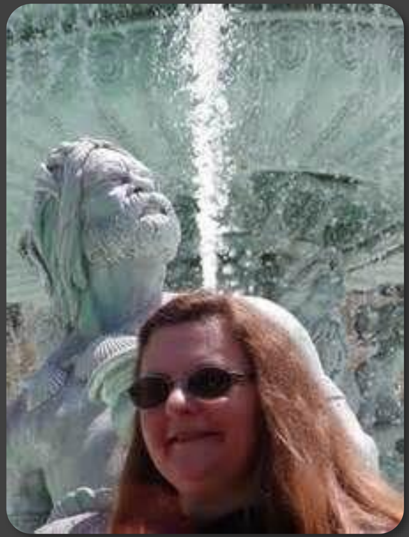
Device



Lighting



Viewpoint



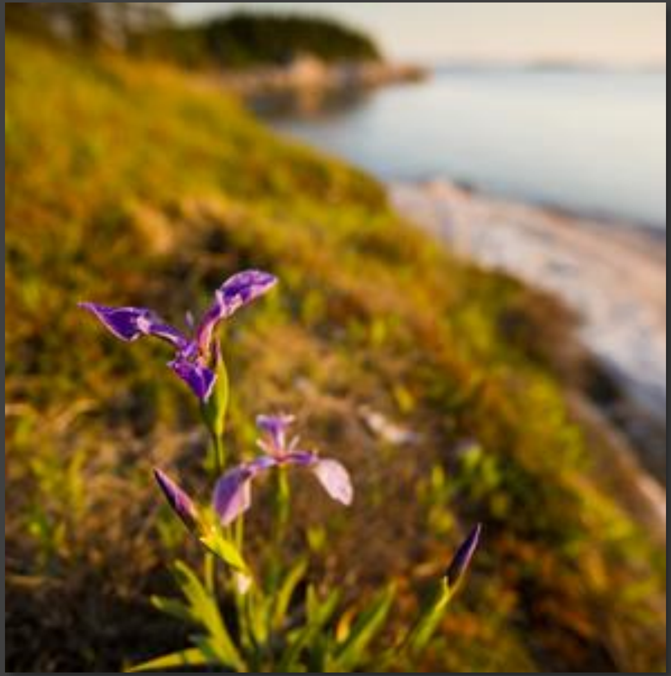
Background

Difficulty in Controlling Imaging Factors

- Numerous factors
- Specialized equipment
  - Inflexible
  - Expensive
- Expertise
- Multiple trials

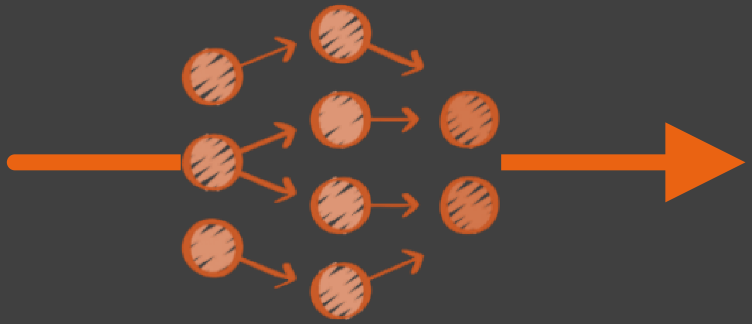
Image credit: DALLE2

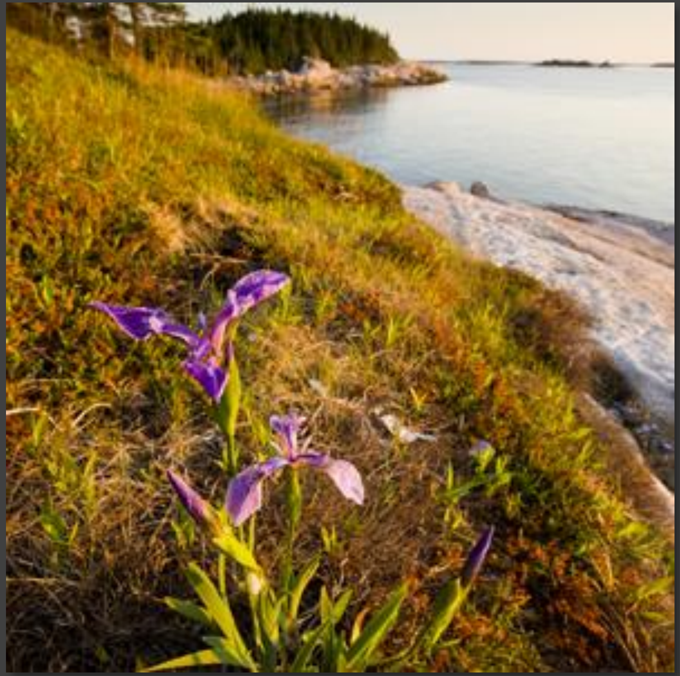
Simple yet Popular DL-based Solution



Undesired
Samplings

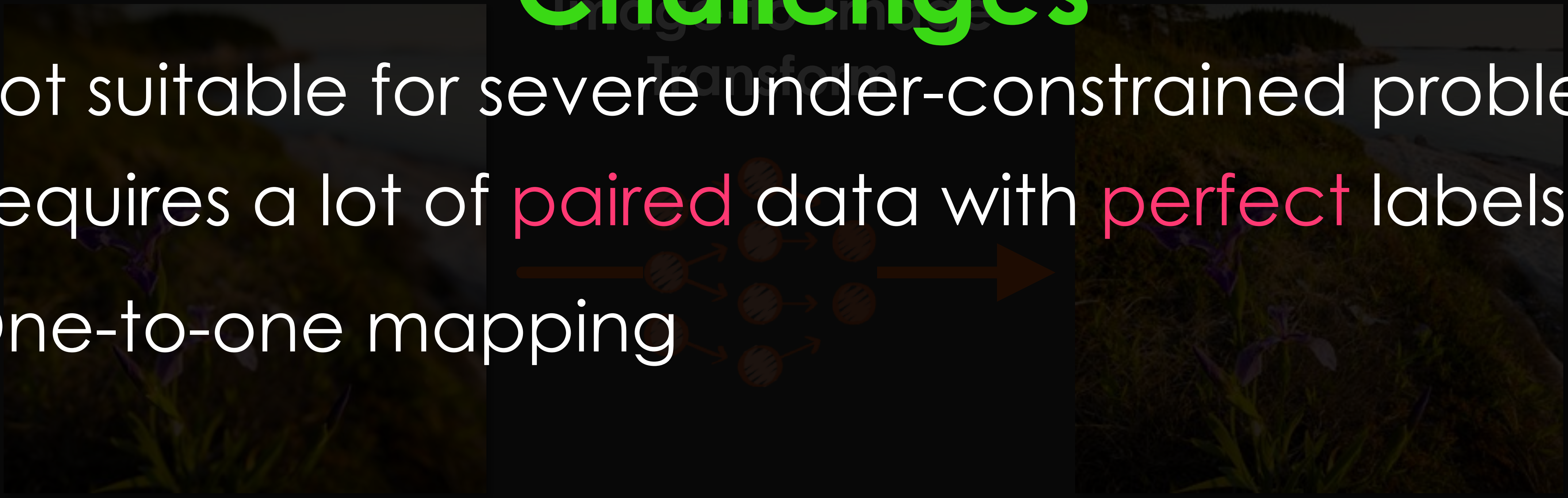
Image-to-Image Transform





Desired
Samplings

Popular Approaches

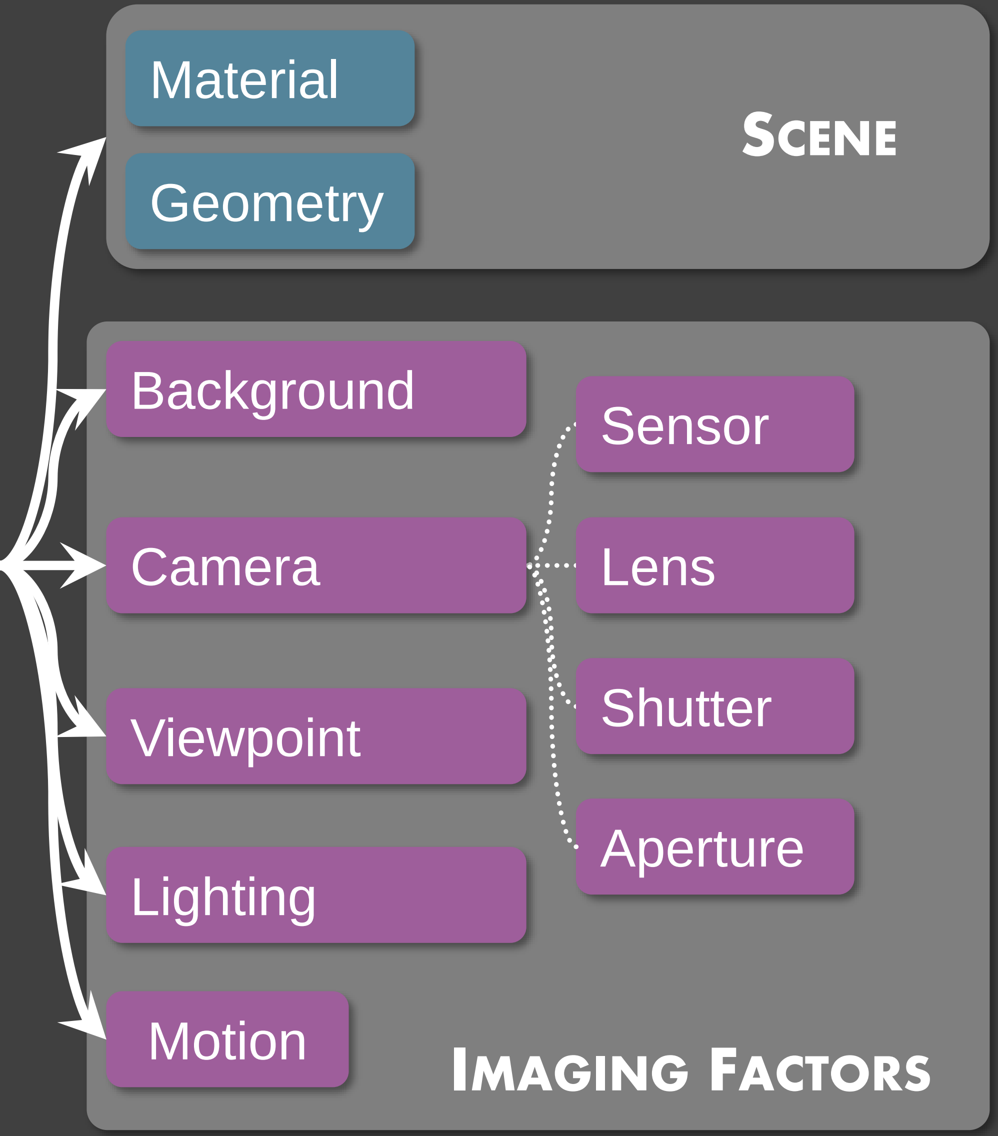


Challenges

- Not suitable for severe under-constrained problems
- Requires a lot of paired data with perfect labels
- One-to-one mapping

Image Factors and Factorization

Undesired Image



Material

Geometry

SCENE

Background

Camera

Viewpoint

Lighting

Motion

Sensor

Lens

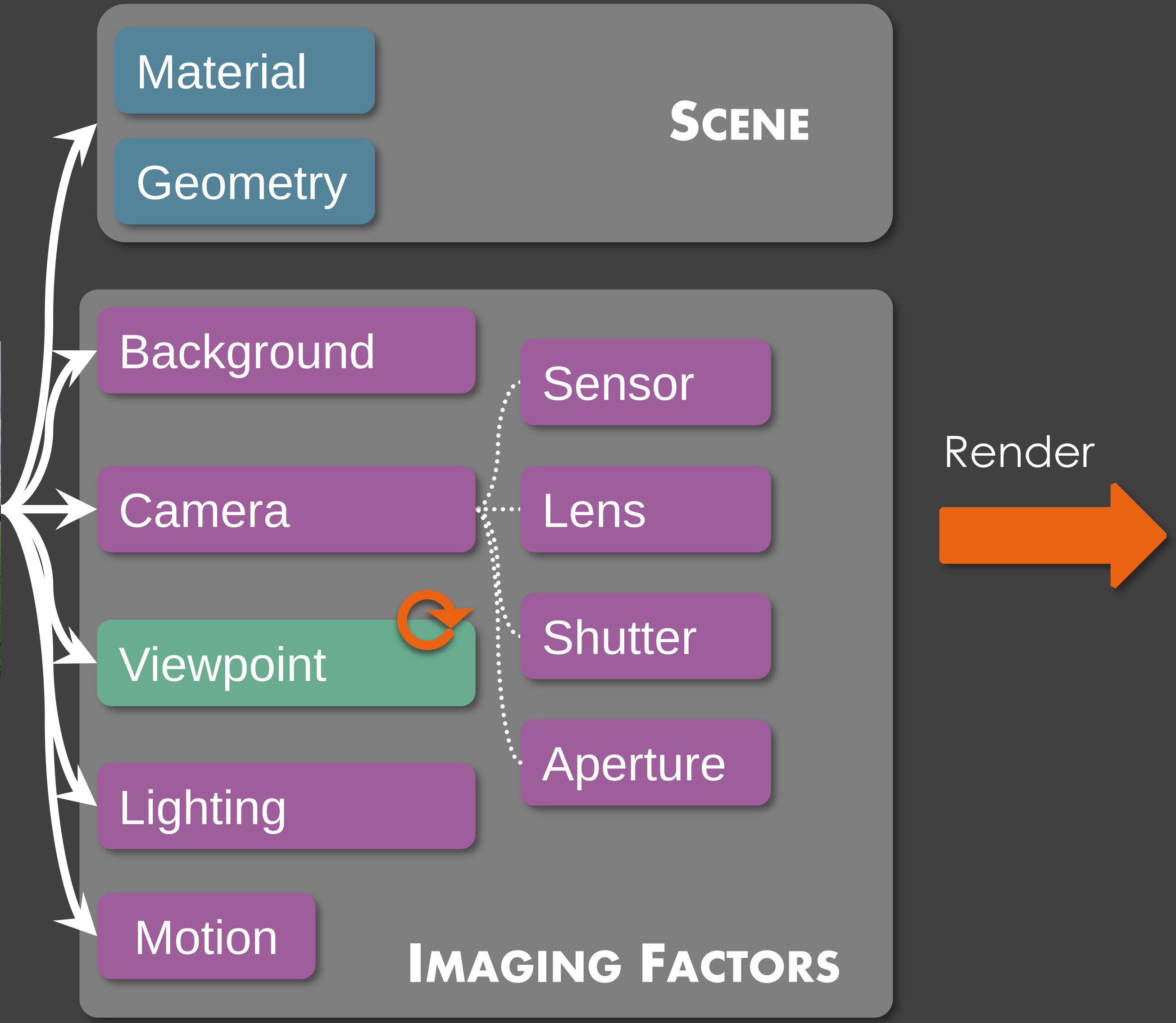
Shutter

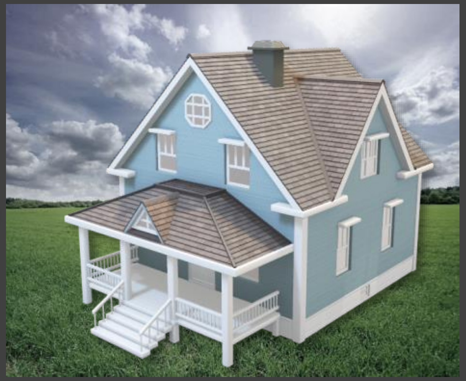
Aperture

IMAGING FACTORS

Image Manipulation

Undesired Image

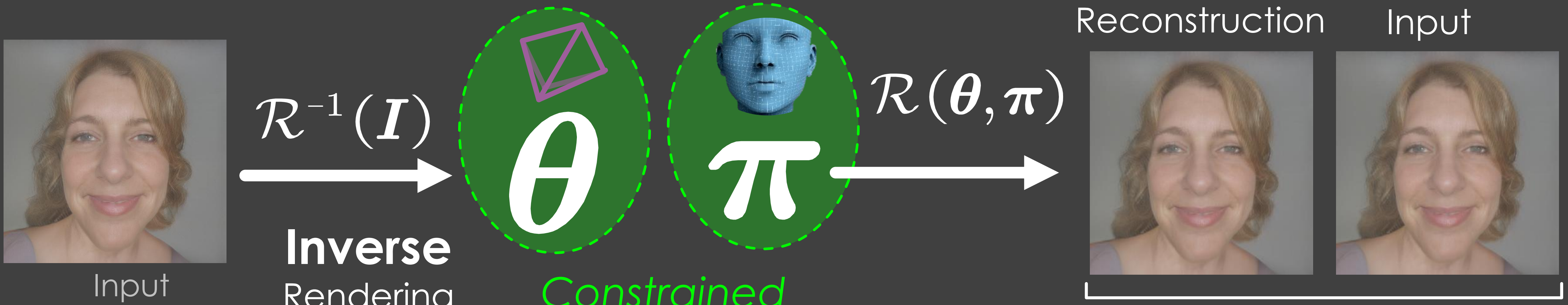




Re-rendering

Harness Pre-trained Generative Models

Optimization-based: no labels required



Input

$\mathcal{R}^{-1}(\boldsymbol{I})$

Inverse Rendering

$\boldsymbol{\theta}$ $\boldsymbol{\pi}$

Constrained

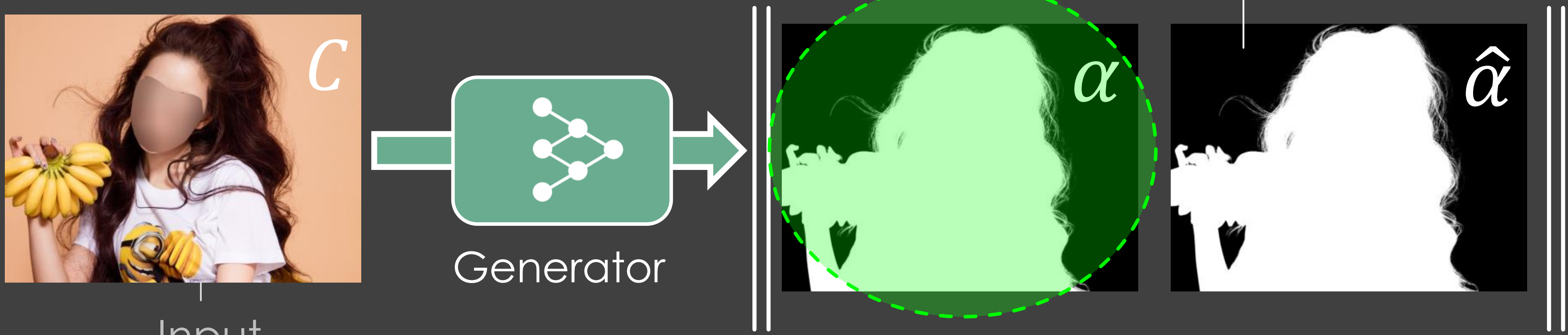
$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\pi})$

Reconstruction Input

+Geometry Loss

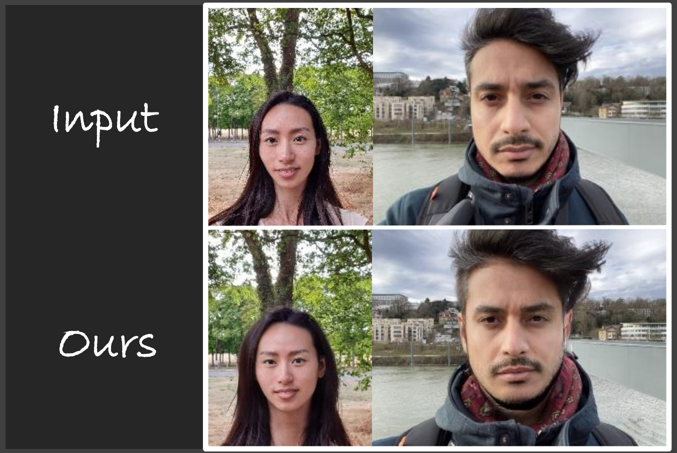
Learning with Labels: imperfect labels

Human Annotations



Input

Viewpoint + Lens



Perspective Distortion Correction

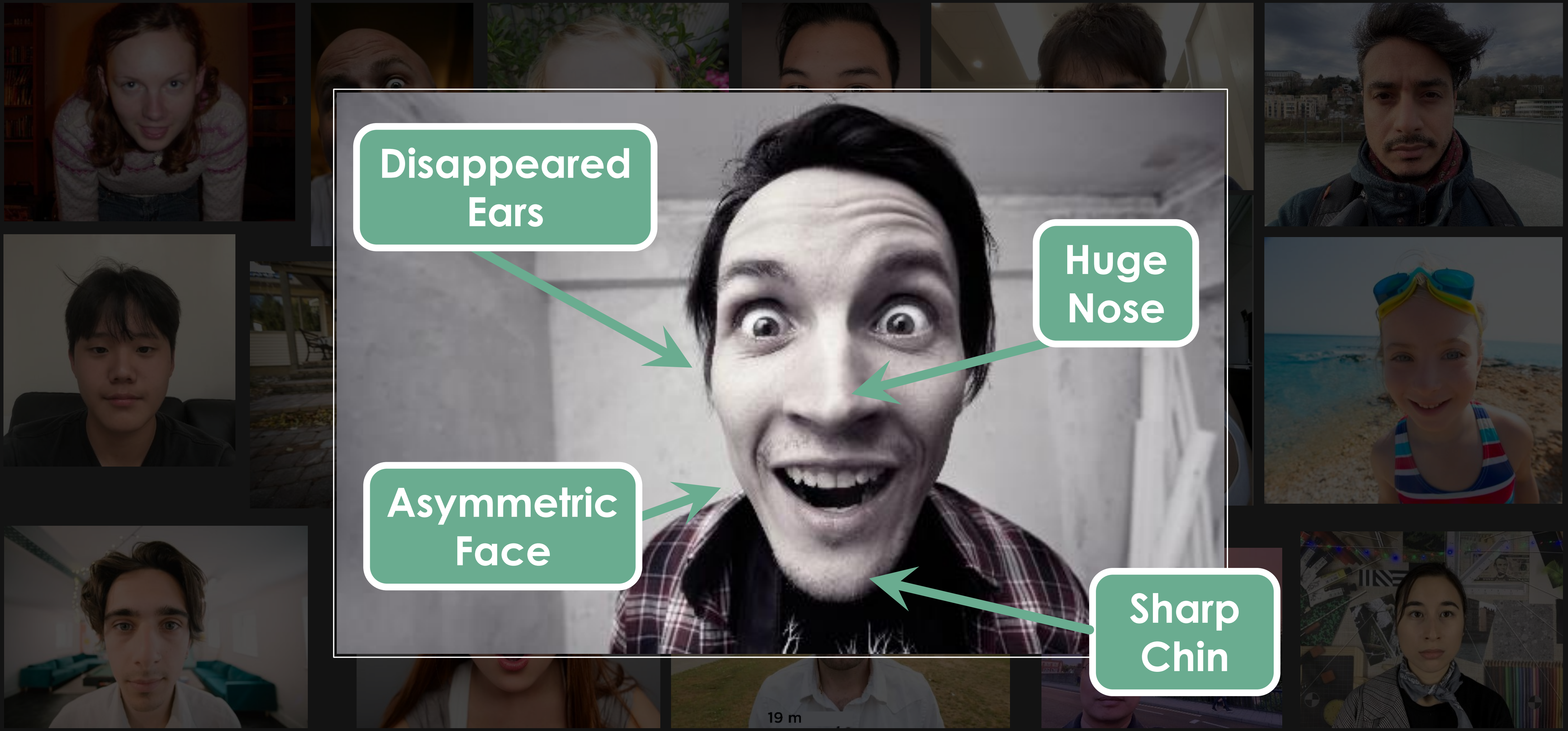
Wang et al, IJCV 2024

Background

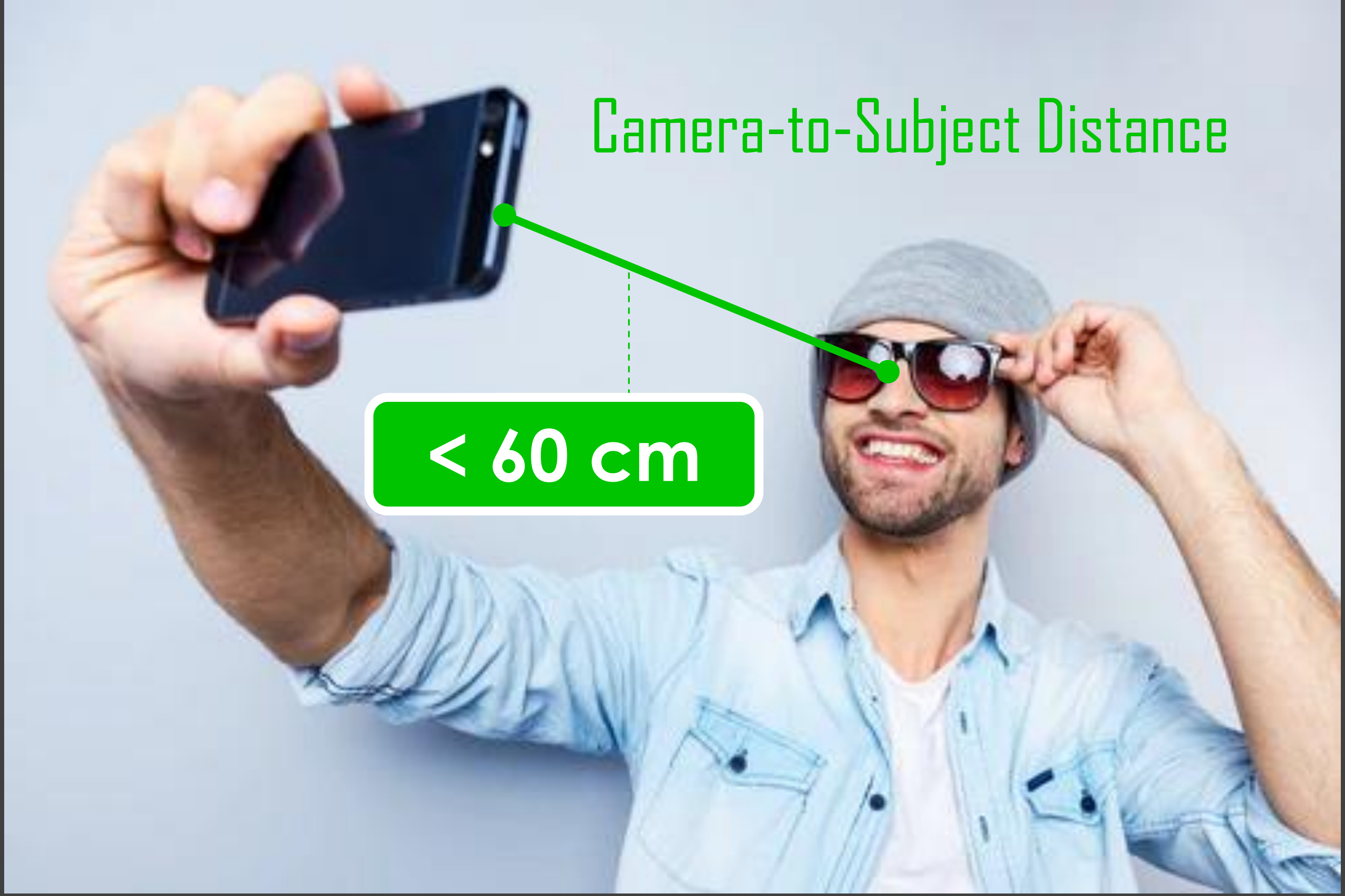
Matting by Generation

Wang et al, SIGGRAPH 2024

CLICK LOVE GROW

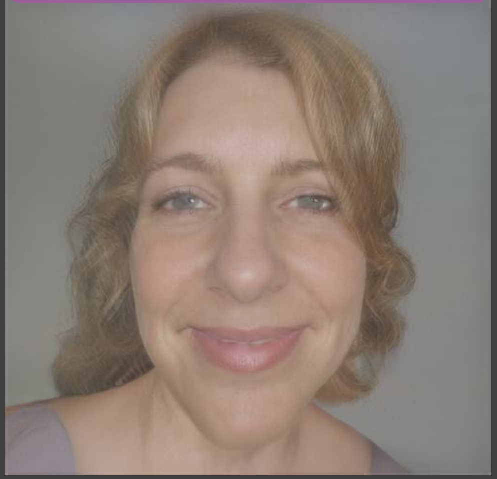

Disappeared Ears
Huge Nose
Asymmetric Face
Sharp Chin
19 m

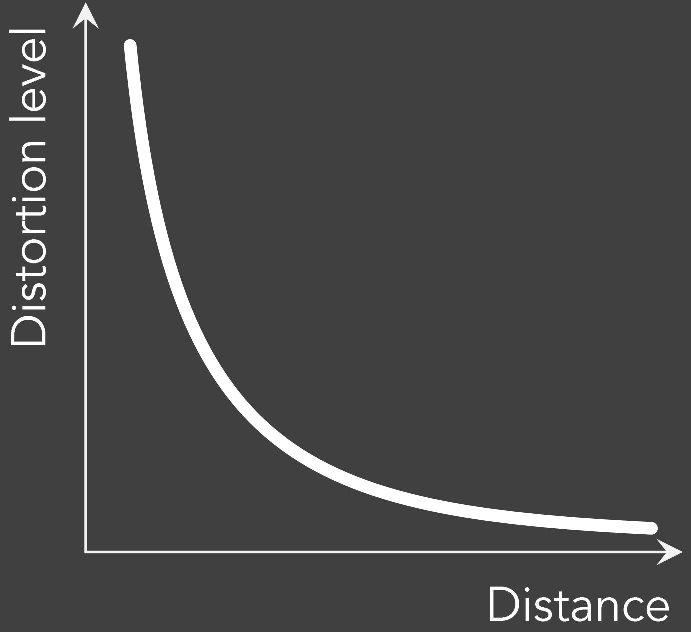
Short Camera-to-Subject Distance



Perspective Projection

Perspective





Depth Variation

Δd

D_{close}

Camera-to-Subject Distance

Weak-perspective Projection

Depth Variation

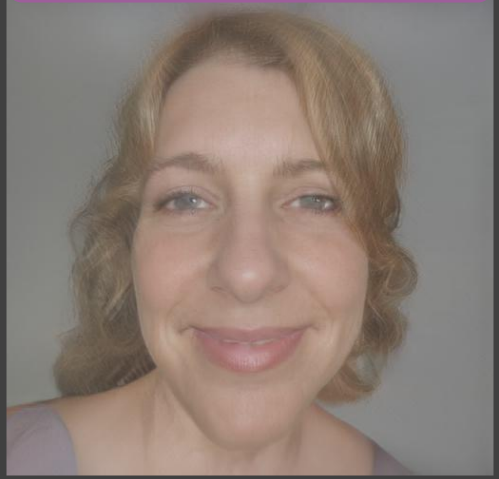
Δd

D_{close}

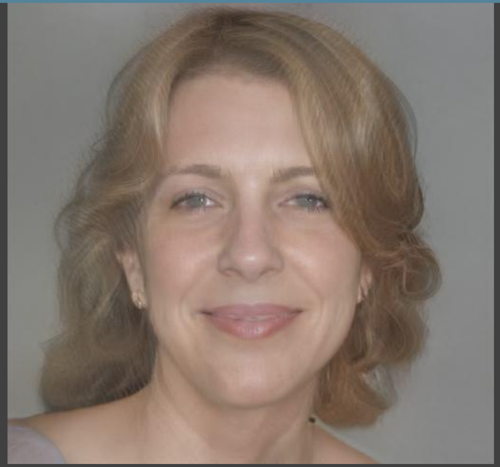
D_{far}

Camera-to-Subject Distance

Perspective



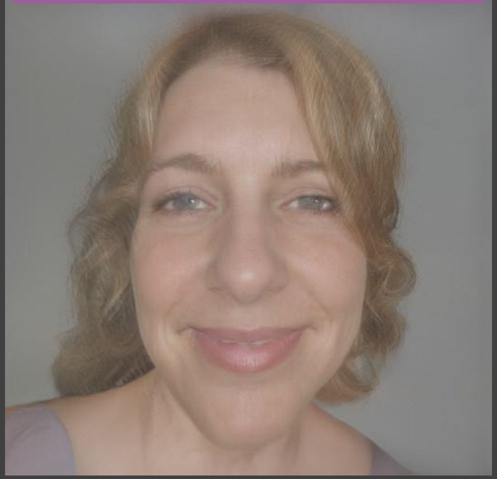
Weak-perspective



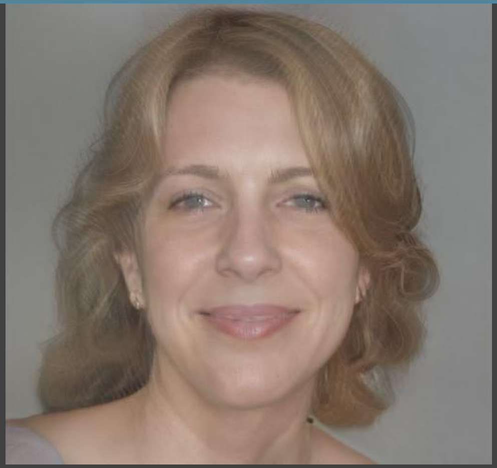
$D_{\text{far}} \gg \Delta d$

Manipulate Viewpoint and Lens

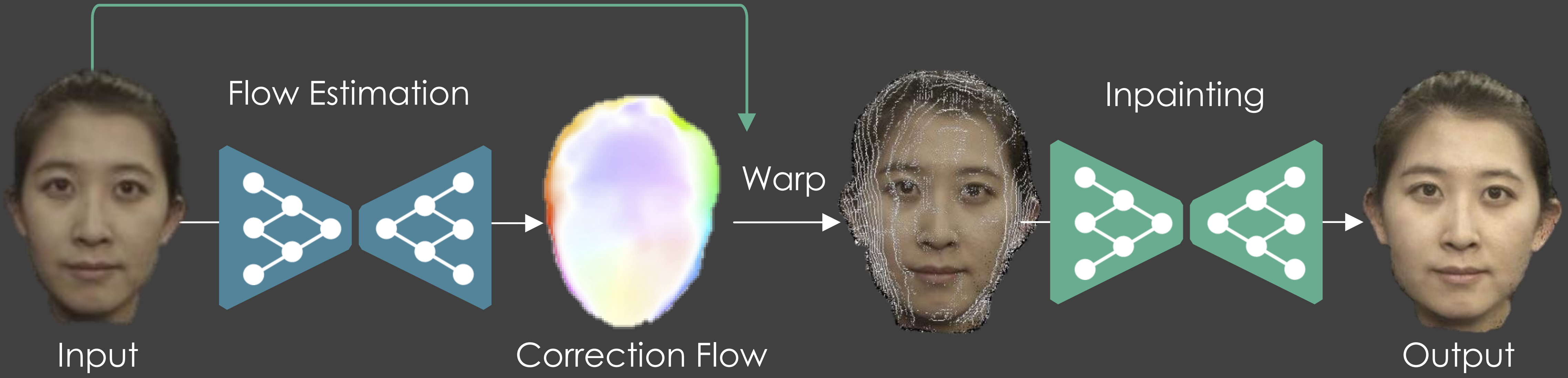
Perspective



Weak-perspective



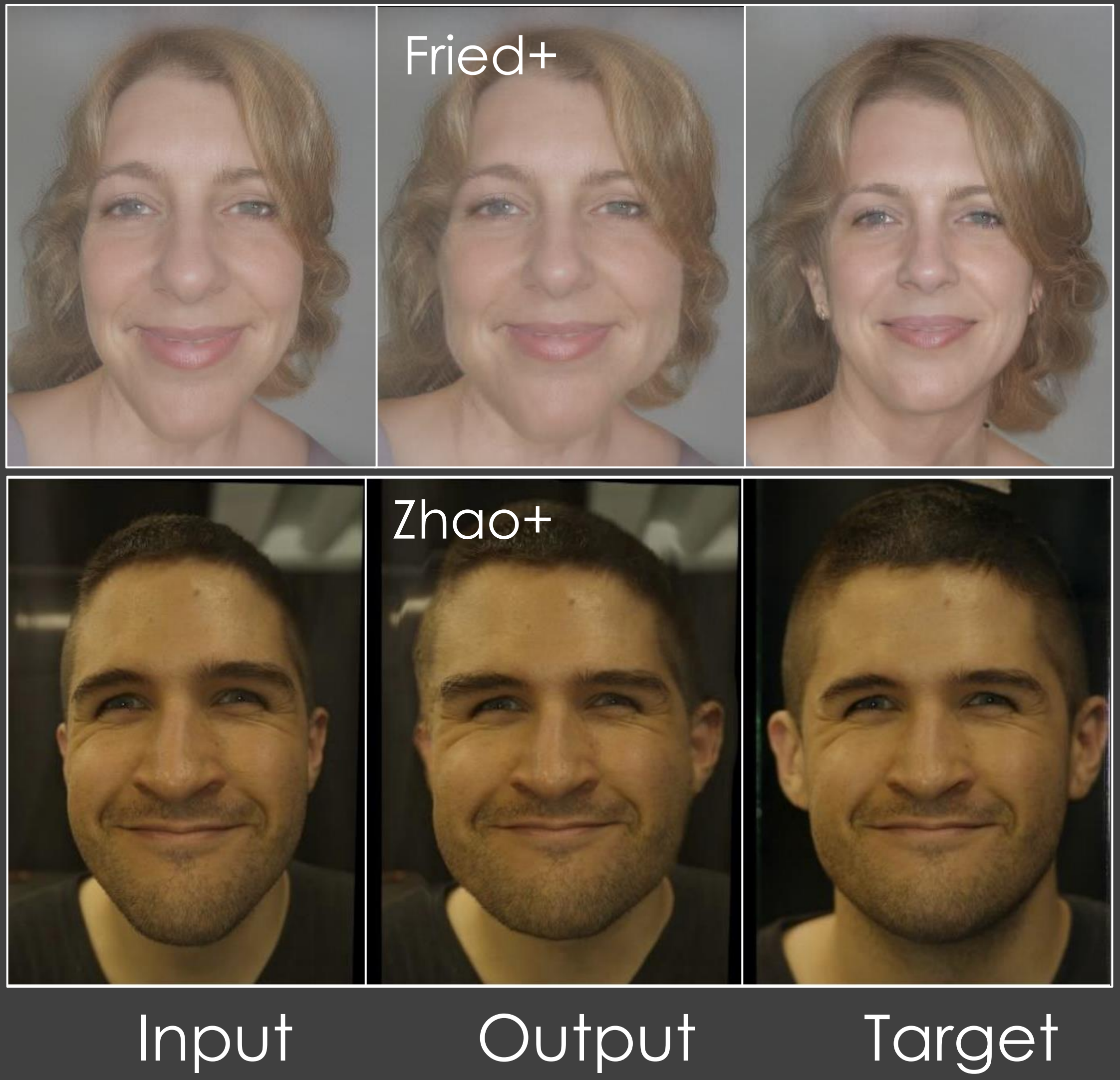
Existing Methods – Warping-based



Fried et al, SIGGRAPH'16

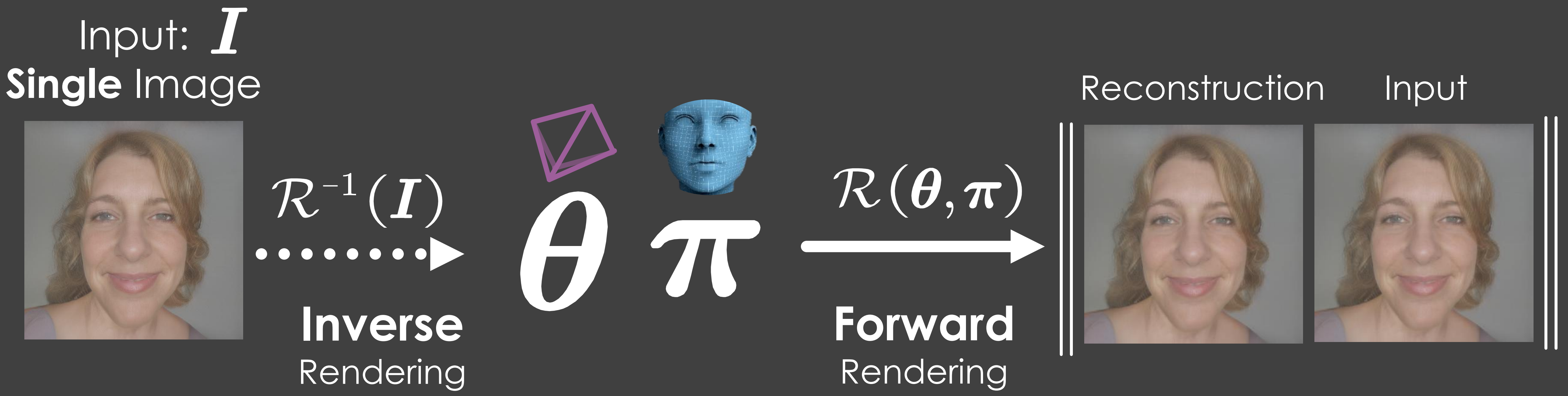
Zhao et al, ICCV'19

Limitations of Existing Methods



- **Flow warping only repeats existing pixels**
  - CANNOT reveal occluded regions
    - Invisible ear, cheek, neck …
  - CANNOT deal with serious distortion
    - When camera-to-face distance is 20–40cm
  - Not 3D-aware
    - Face shape is flawed
- **Learning-based method (Zhao+) is worse**
  - Require a lot of training data
  - Hard to generalize
  - CANNOT continuously change

Optimization-based Factorization



Input: $\boldsymbol{I}$

Single Image

$\mathcal{R}^{-1}(\boldsymbol{I})$

Inverse Rendering

$\boldsymbol{\theta}$ $\boldsymbol{\pi}$

$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\pi})$

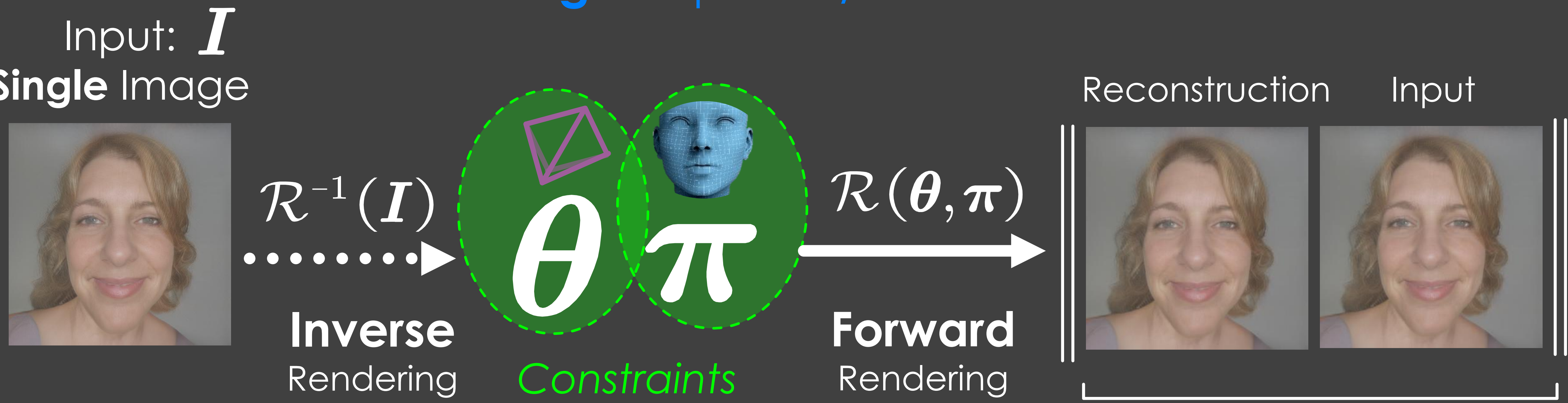
Forward Rendering

Reconstruction

Input

Optimization-based Factorization

Challenge: ill-posed/unconstrained



Input: $\boldsymbol{I}$

Single Image

$\mathcal{R}^{-1}(\boldsymbol{I})$

Inverse Rendering

$\boldsymbol{\theta}$ $\boldsymbol{\pi}$

Constraints

$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\pi})$

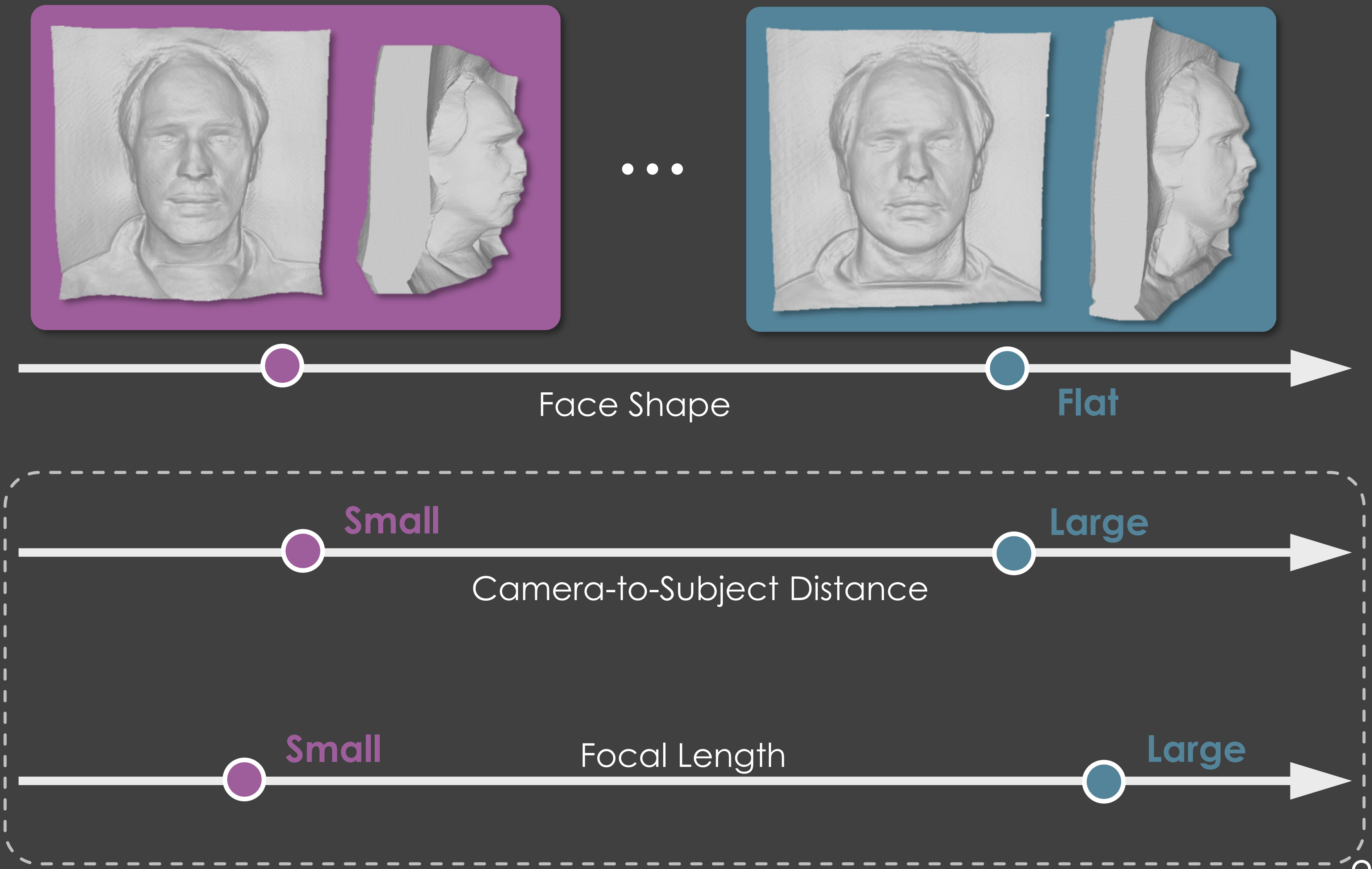
Forward Rendering

Reconstruction Input

+Geometry Loss

Ambiguity of Parameters

Many **combinations** resemble input image



…

Face Shape

Flat

Small

Large

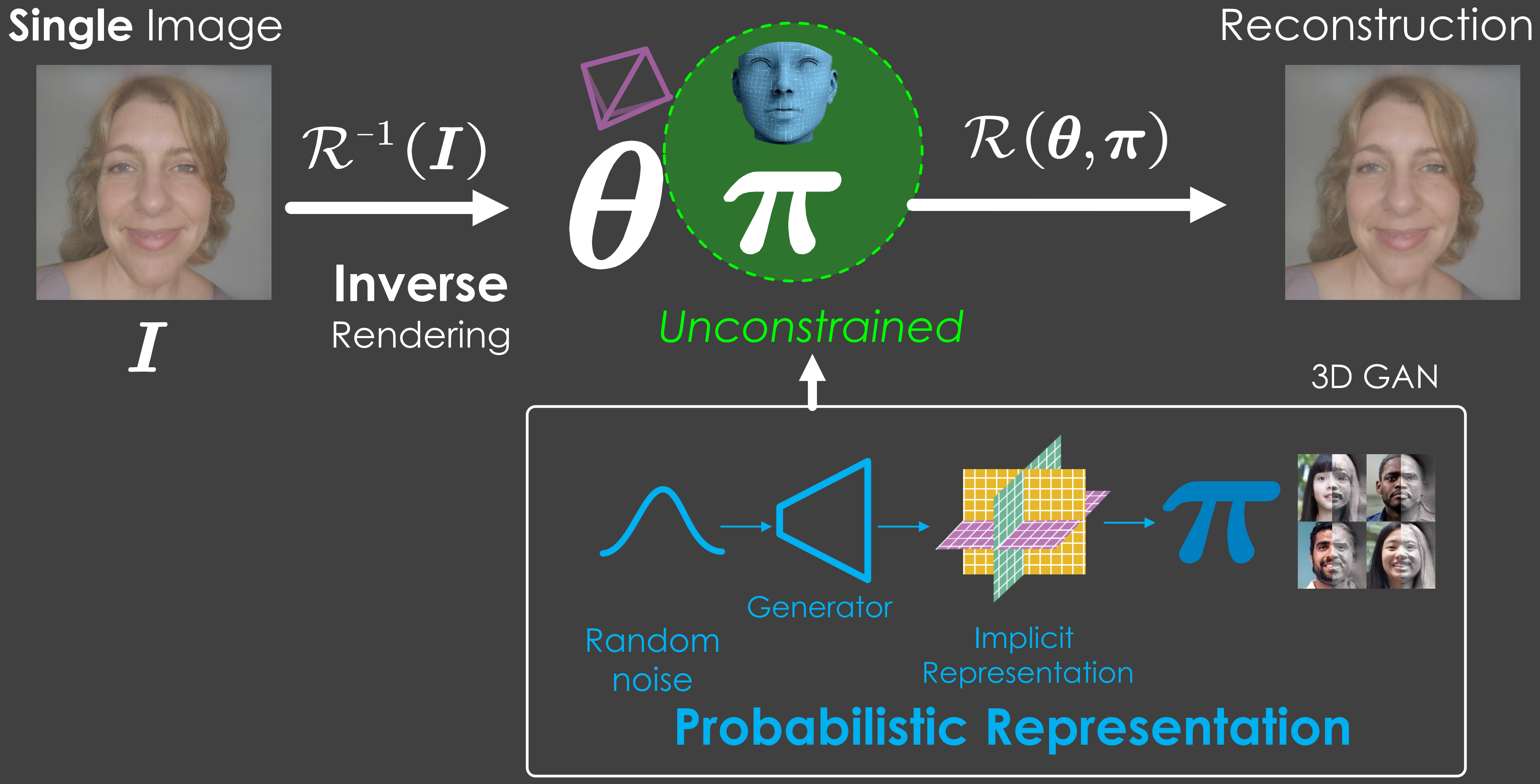
Camera-to-Subject Distance

Small

Focal Length

Large

3D GAN Prior as Face Constraint



Single Image

$\boldsymbol{I}$

$\mathcal{R}^{-1}(\boldsymbol{I})$

Inverse Rendering

$\boldsymbol{\theta}$ $\boldsymbol{\pi}$

Unconstrained

$\mathcal{R}(\boldsymbol{\theta},\boldsymbol{\pi})$

Reconstruction

3D GAN

Random noise

Generator

Implicit Representation

$\boldsymbol{\pi}$

Probabilistic Representation

Camera Regularization (CR)

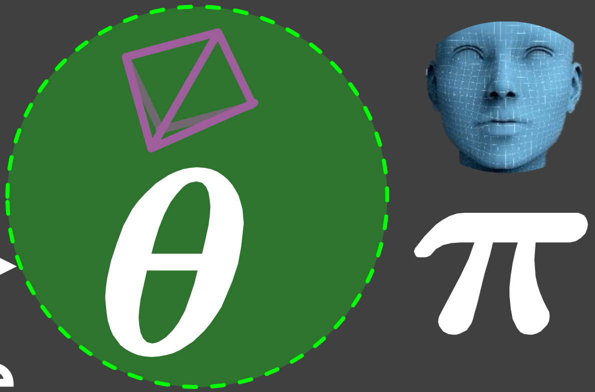
Input: $\boldsymbol{I}$

Single Image

$\mathcal{R}^{-1}(\boldsymbol{I})$

Inverse

Rendering



Unconstrained

3 Strategies

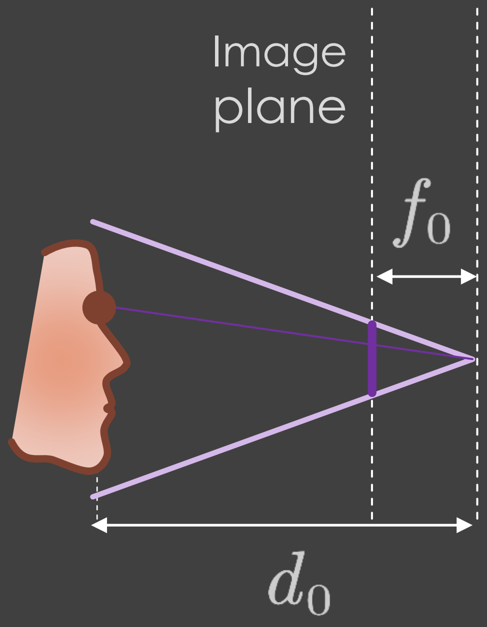
$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\pi})$

Reconstruction

CR 1: Focal Length Re-parameterization

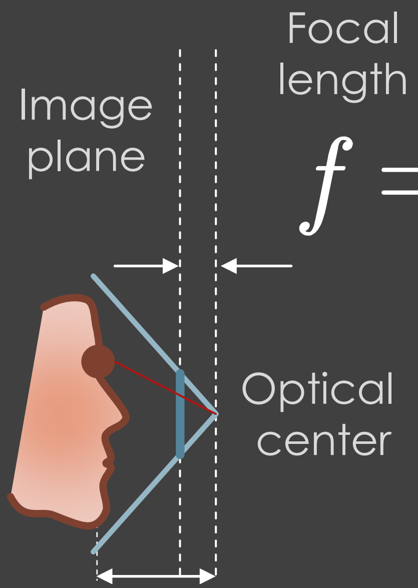
- Focal Length

(simplified approximation)



Camera-to-Subject Distance

$$\Delta\mathbf{t} = \begin{pmatrix} \Delta t_x \\ \Delta t_y \\ \Delta t_z \end{pmatrix}$$

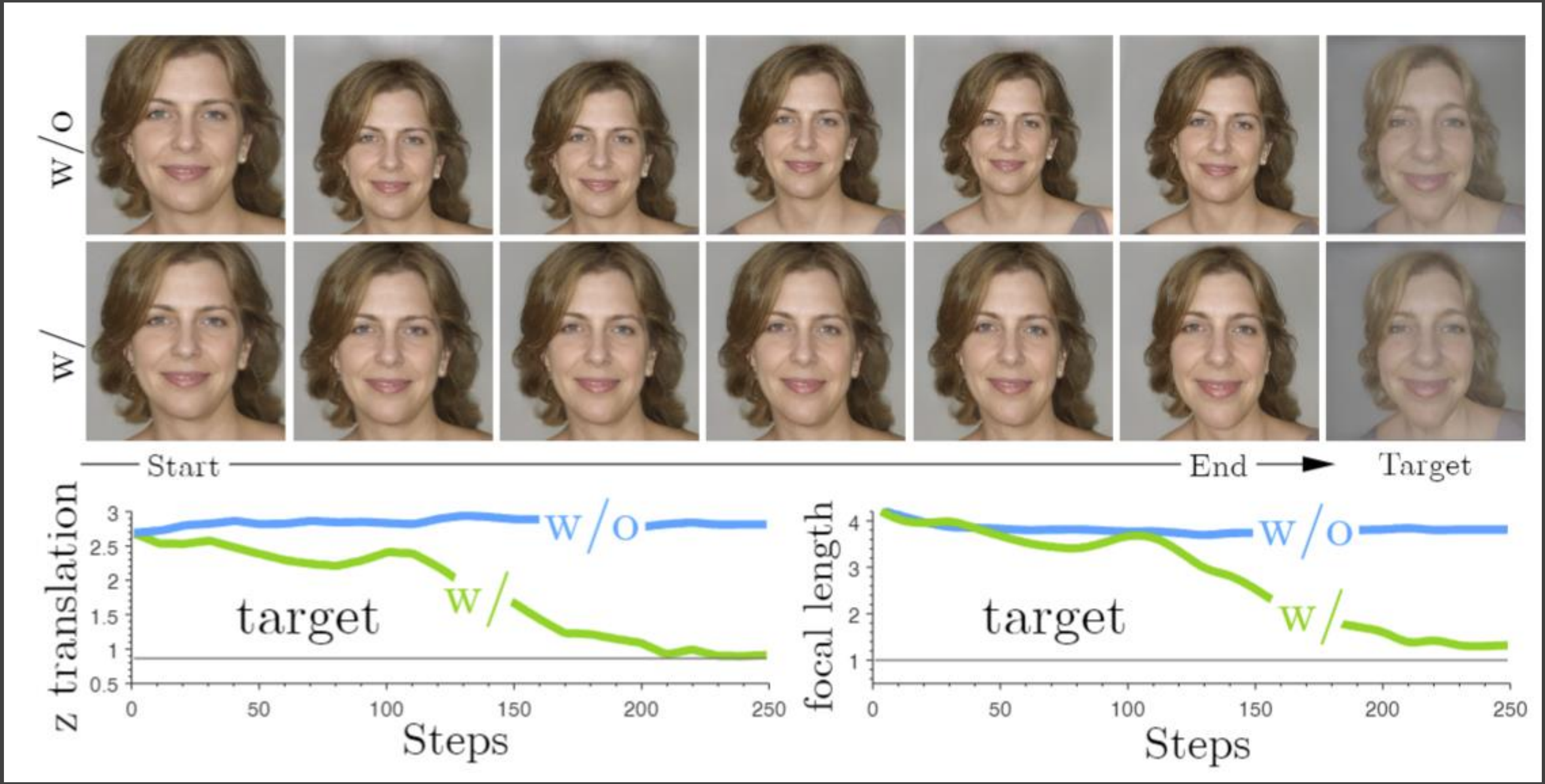


$$f = \frac{d}{d_0} f_0$$

$$d = d_0 + \Delta t_z$$

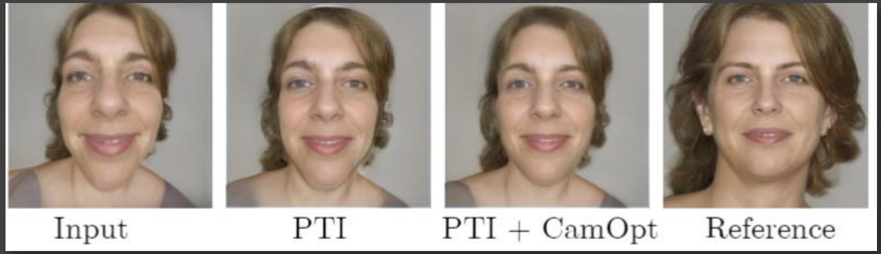
Motivation: Reduce unknown parameters and decouple

CR 1: Focal Length Re-parameterization

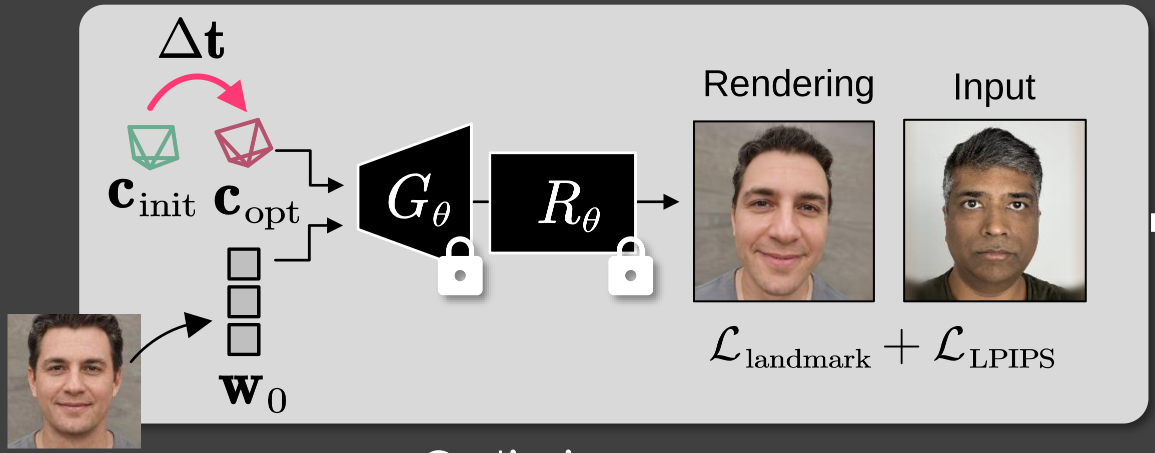


CR 2: Optimization Scheduling

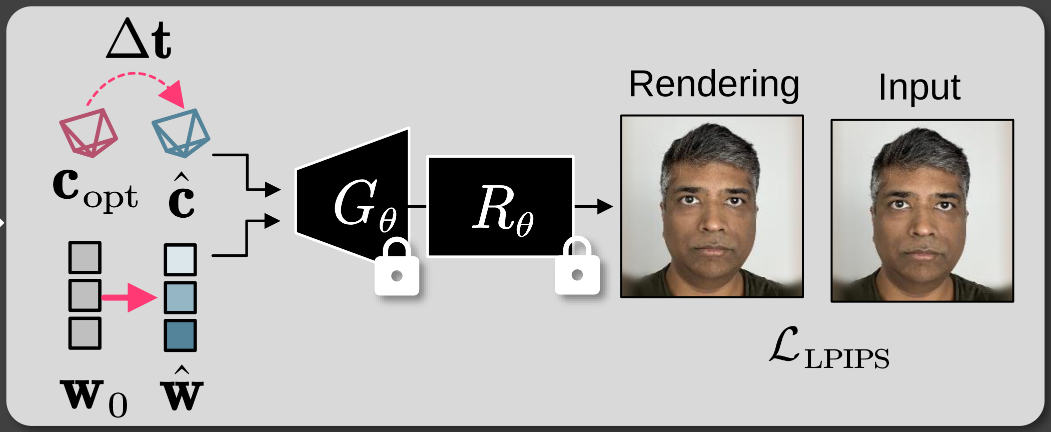
Motivation: Face is **easier** to fall into **sub-optimum** than camera



Optimization   Frozen


Δt
$\mathbf{c}_{\text{init}}$
$\mathbf{c}_{\text{opt}}$
$\mathbf{w}_0$
G_θ
R_θ
Rendering
Input
$\mathcal{L}_{\text{landmark}} + \mathcal{L}_{\text{LPIPS}}$

Optimize camera



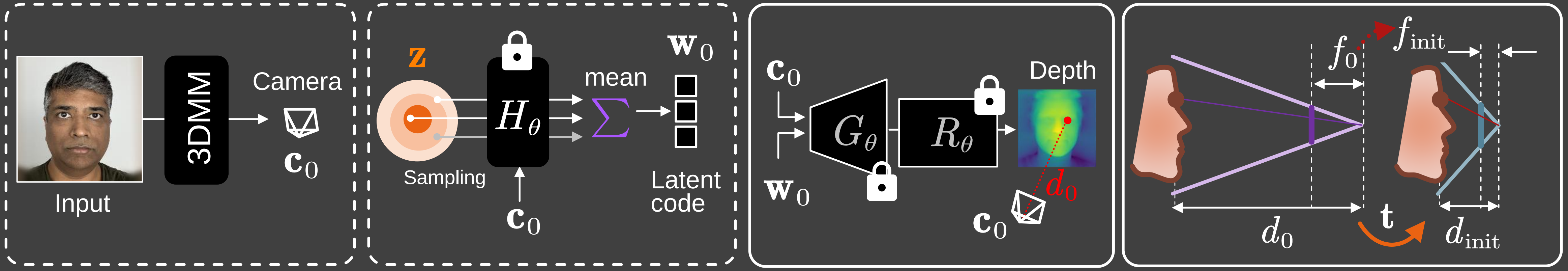
Optimize face, refine camera

CR 3: Better Initialization

Start from a close-up camera position

Original initialization

Re-parameterization



Get init. cam $\mathbf{c}_0$

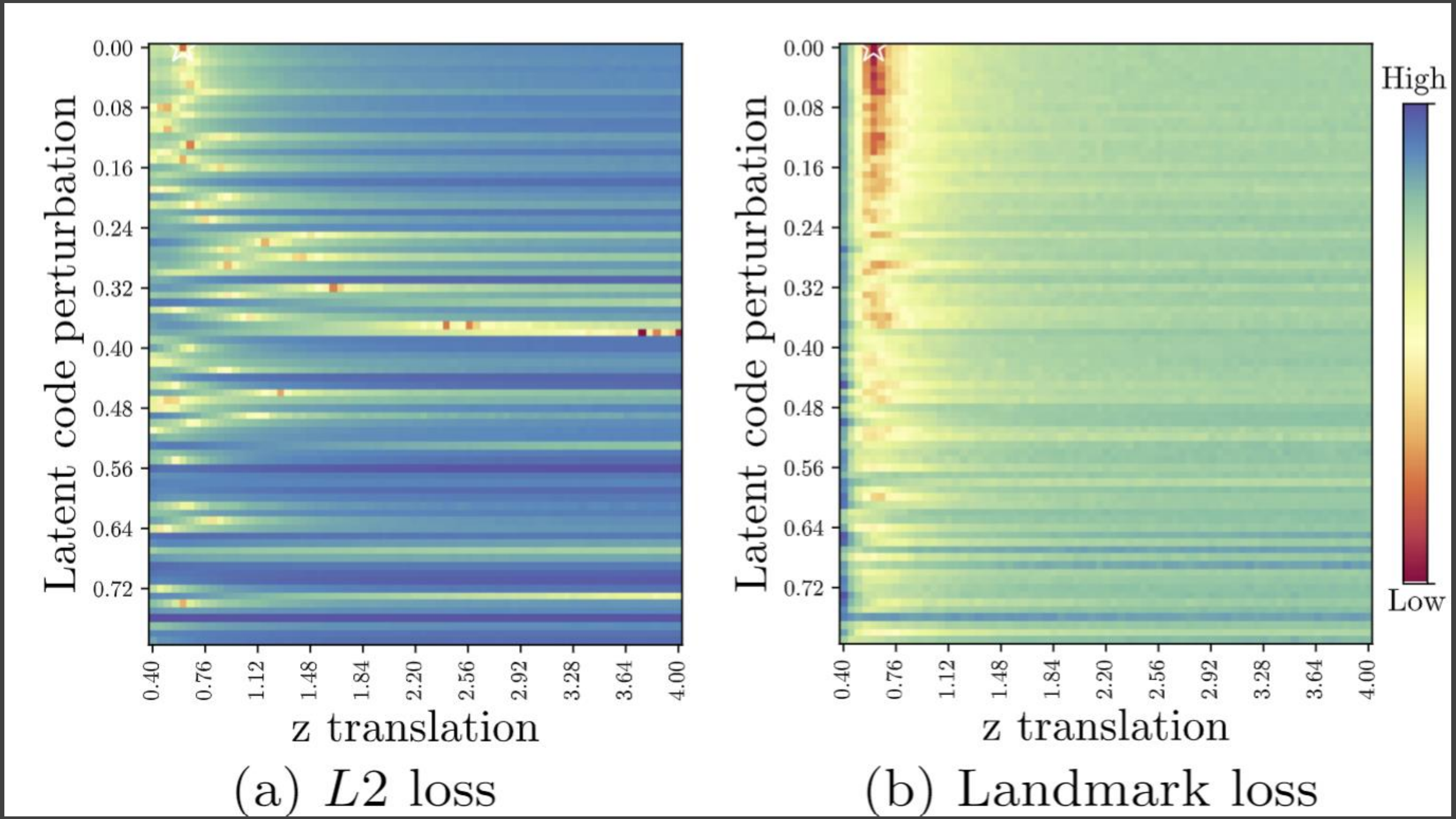
Get init. face para. $\mathbf{w}_0$

Get eyes' distance d_0

Modify $\mathbf{c}_0 \rightarrow \mathbf{c}_{\text{init}}$

Ambiguity Caused by Loss

Pixel loss is **very sensitive** to pixel change



(a) $L2$ loss

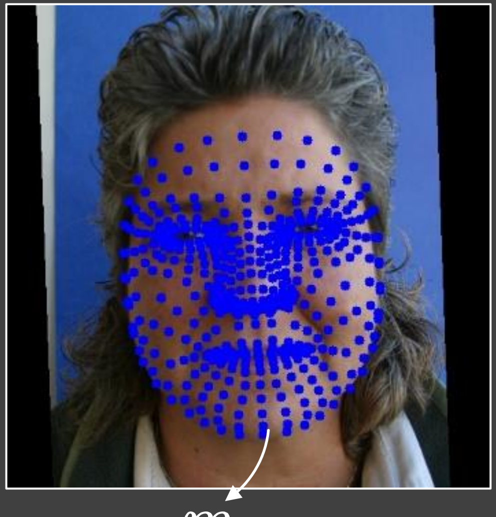
(b) Landmark loss

Geometric Regularization

Uncertainty-based Loss

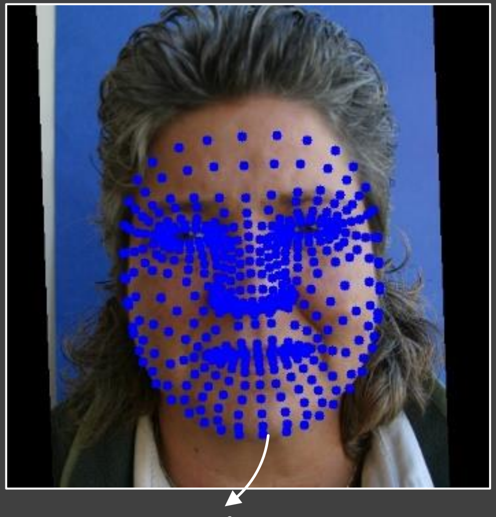
$$\sum_{i=1}^{\|\mathcal{M}\|}\left(\underbrace{\log\left(\sigma_i^2\right)}_{\textbf{Uncertainty}\text{ term}} + \frac{\|m_i - m_i'\|_2^2}{2\sigma_i^2}\right)$$

prediction



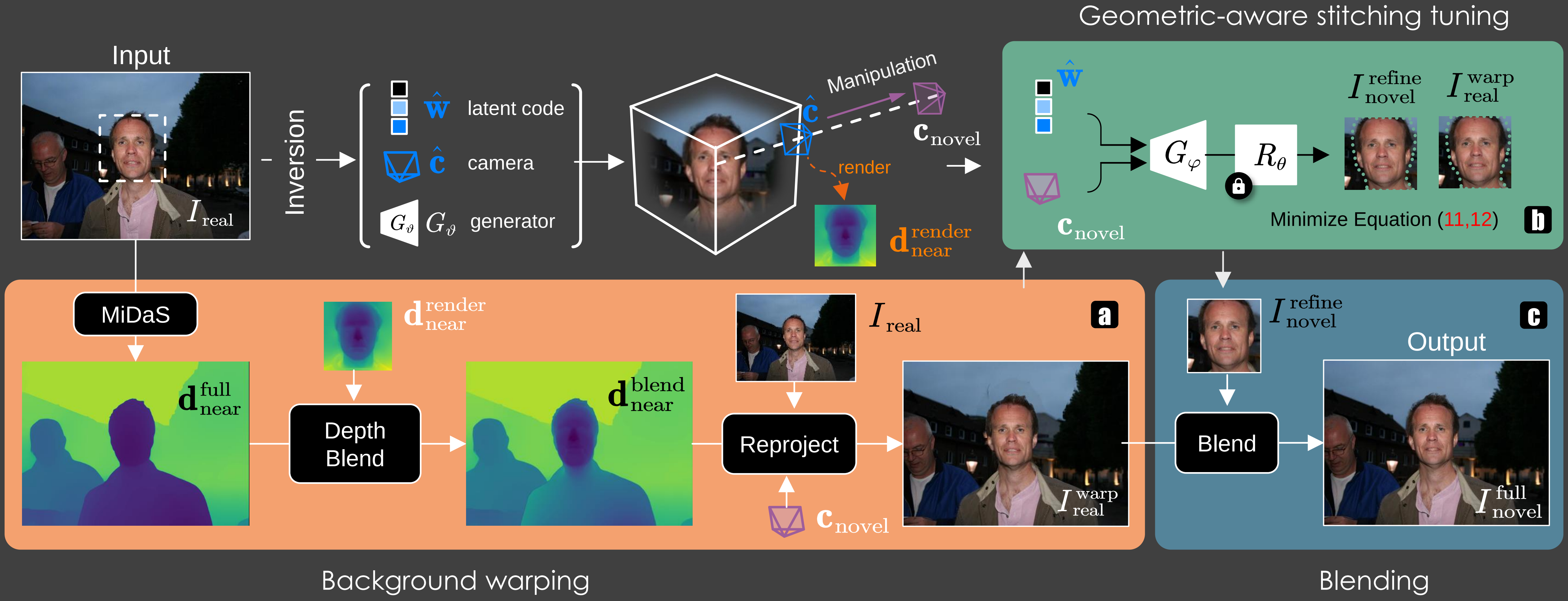
m_i

input

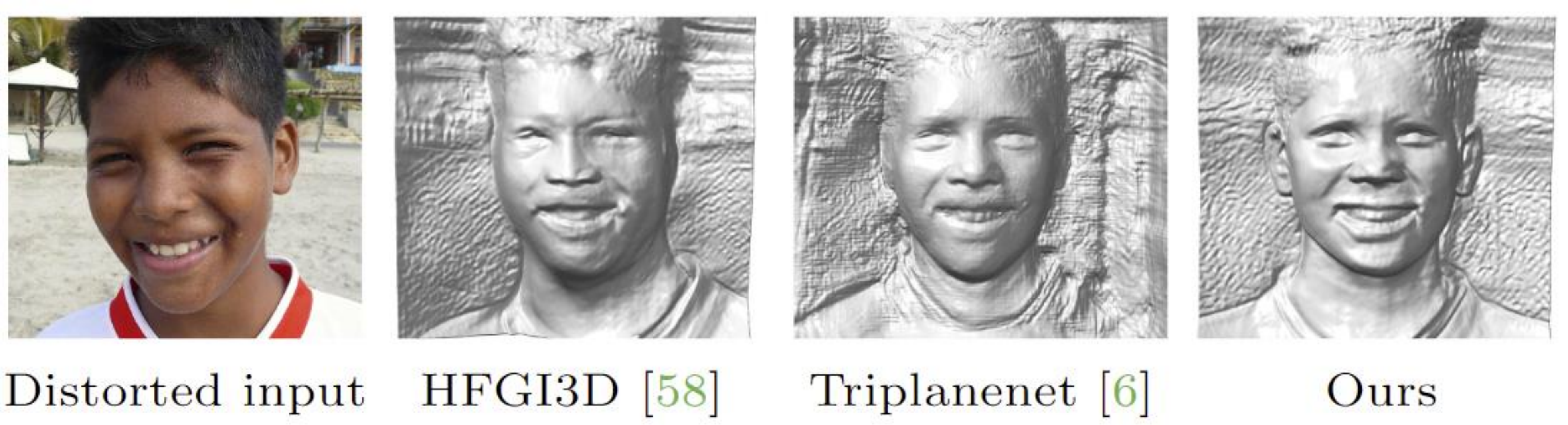


m_i'

Extensions for Full-frame Image


Geometric-aware stitching tuning
Input
I_{real}
Inversion
$\hat{\mathbf{w}}$ latent code
$\hat{\mathbf{c}}$ camera
G_ϑ G_ϑ generator
$\hat{\mathbf{c}}$
Manipulation
$\mathbf{c}_{\text{novel}}$
render
$\mathbf{d}^{\text{render}}_{\text{near}}$
$\hat{\mathbf{w}}$
G_φ
R_θ
$I^{\text{refine}}_{\text{novel}}$
$I^{\text{warp}}_{\text{real}}$
$\mathbf{c}_{\text{novel}}$
Minimize Equation (11,12)
b
MiDaS
$\mathbf{d}^{\text{full}}_{\text{near}}$
$\mathbf{d}^{\text{render}}_{\text{near}}$
Depth Blend
$\mathbf{d}^{\text{blend}}_{\text{near}}$
I_{real}
Reproject
$\mathbf{c}_{\text{novel}}$
$I^{\text{warp}}_{\text{real}}$
a
$I^{\text{refine}}_{\text{novel}}$
c
Output
Blend
$I^{\text{full}}_{\text{novel}}$
Background warping
Blending

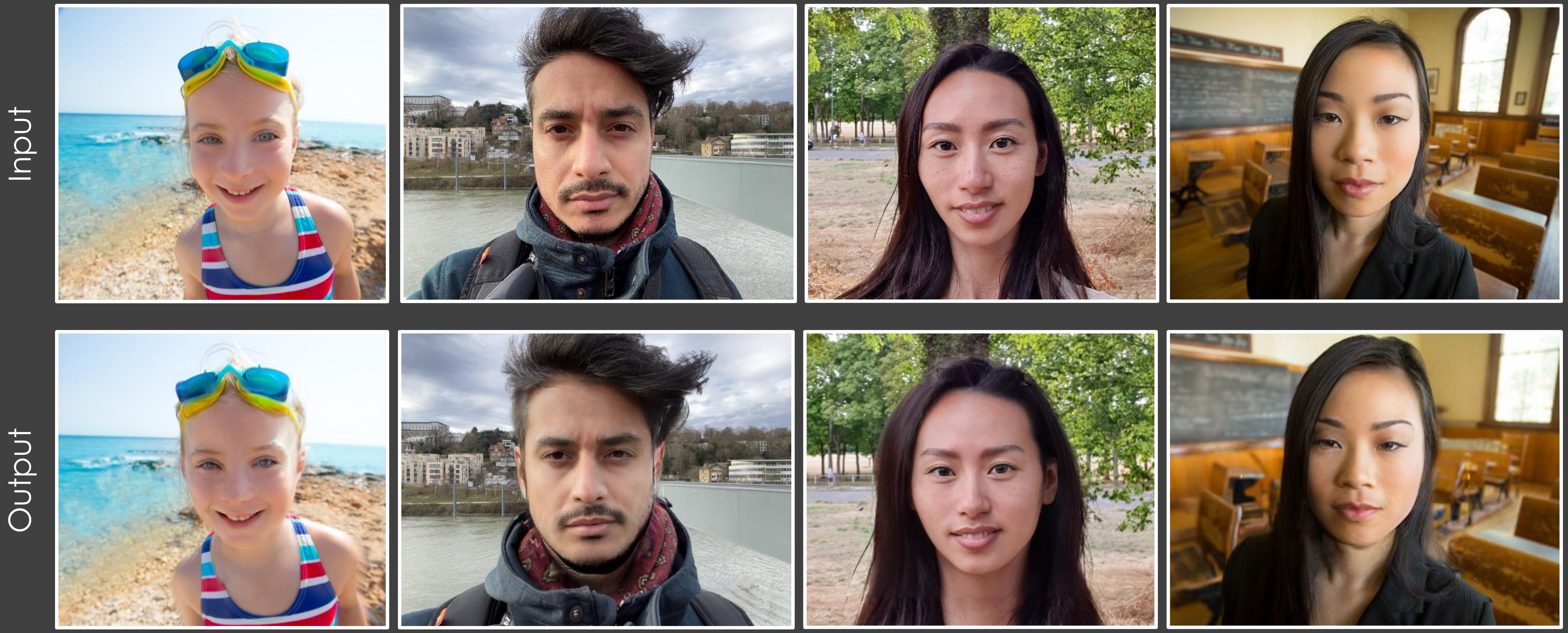
Results – Mesh



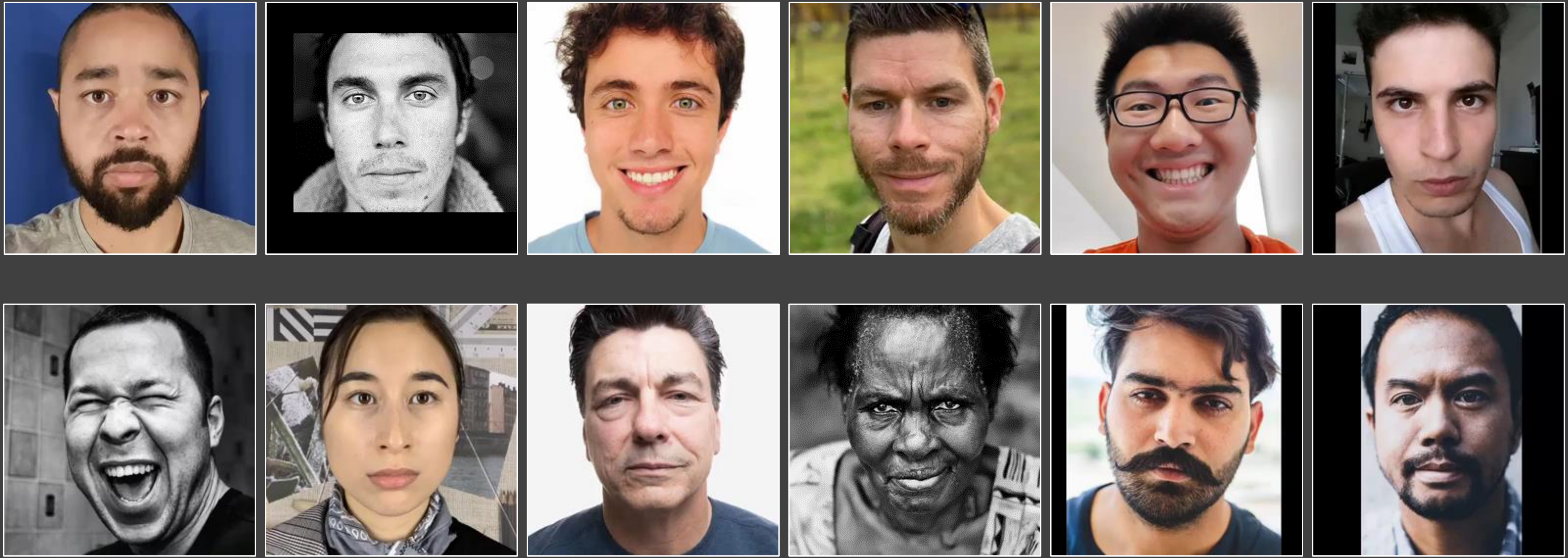
Distorted input HFGI3D [58] Triplanenet [6] Ours

Other GAN inversion methods

Results



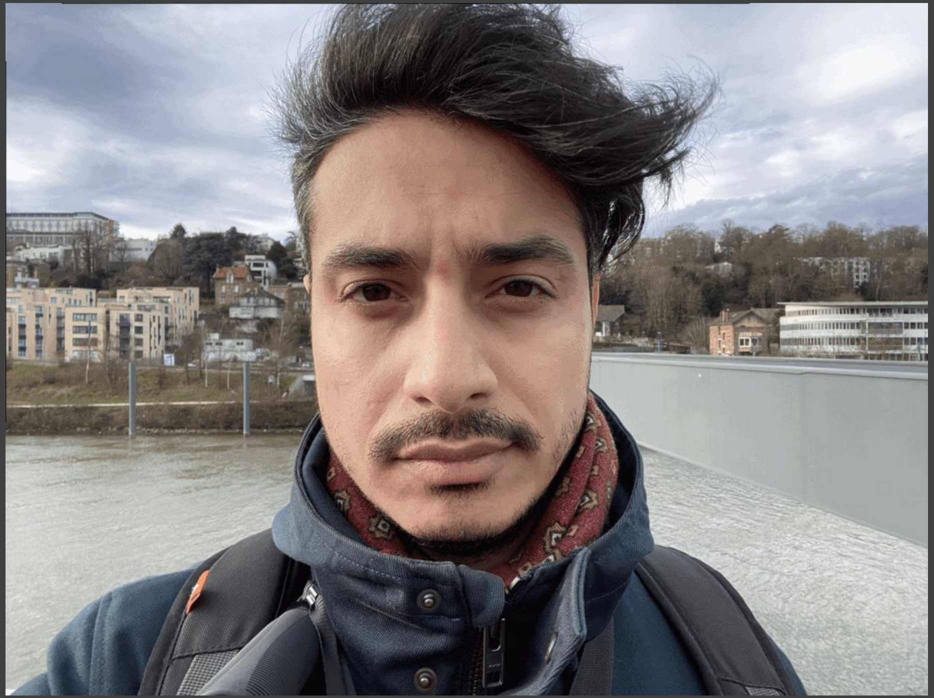
Results – Continuous Manipulation



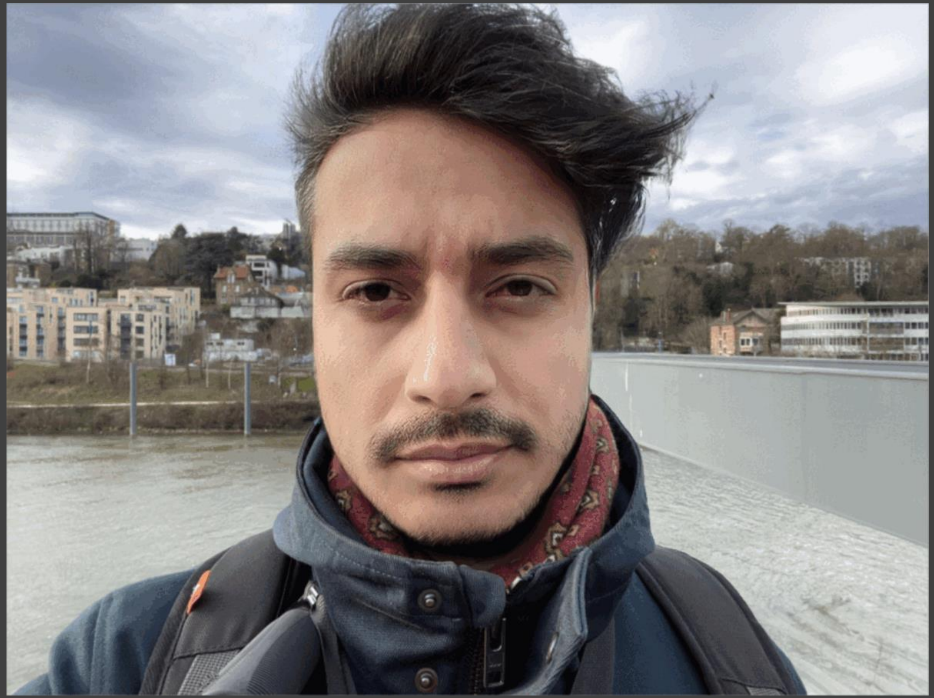
Results – Comparison

Stretch-like 👎

3D **geometric consistent** 👍

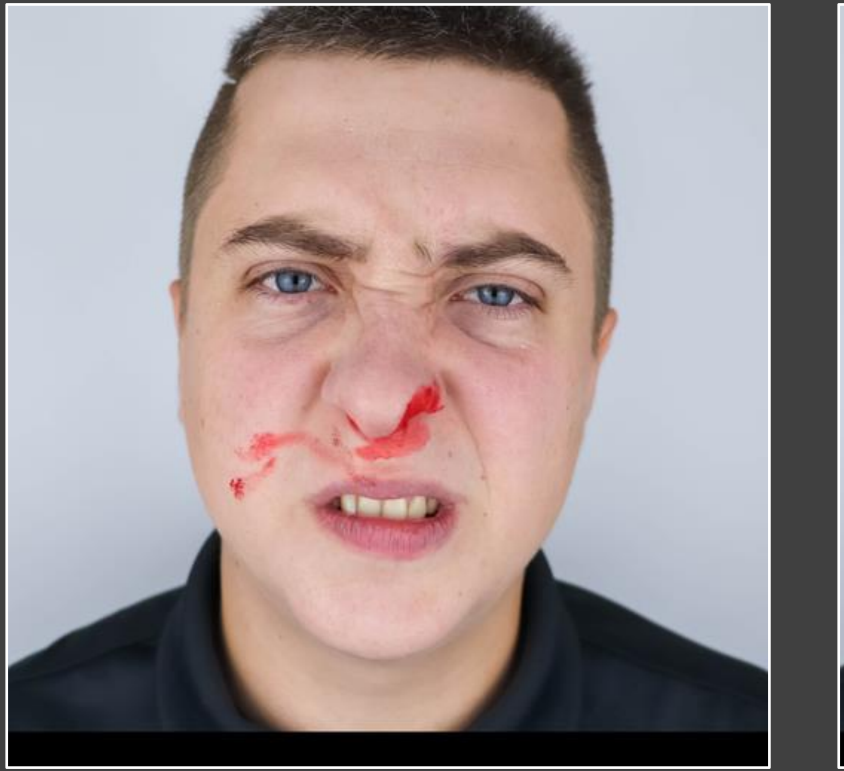


Fried et al, SIGGRAPH’16

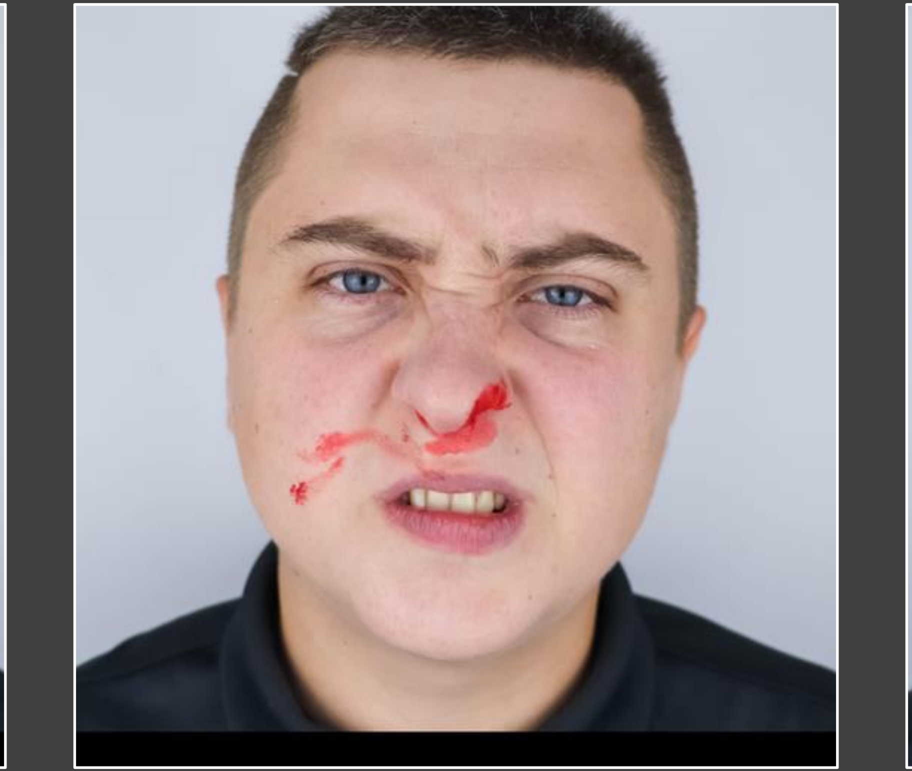


Ours

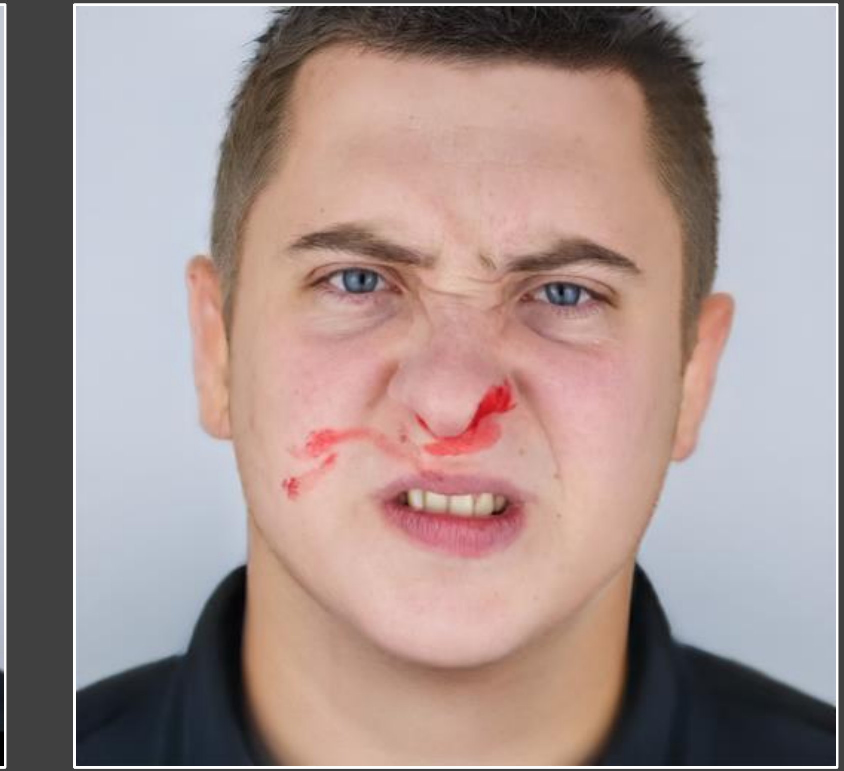
Results – Comparison



Input

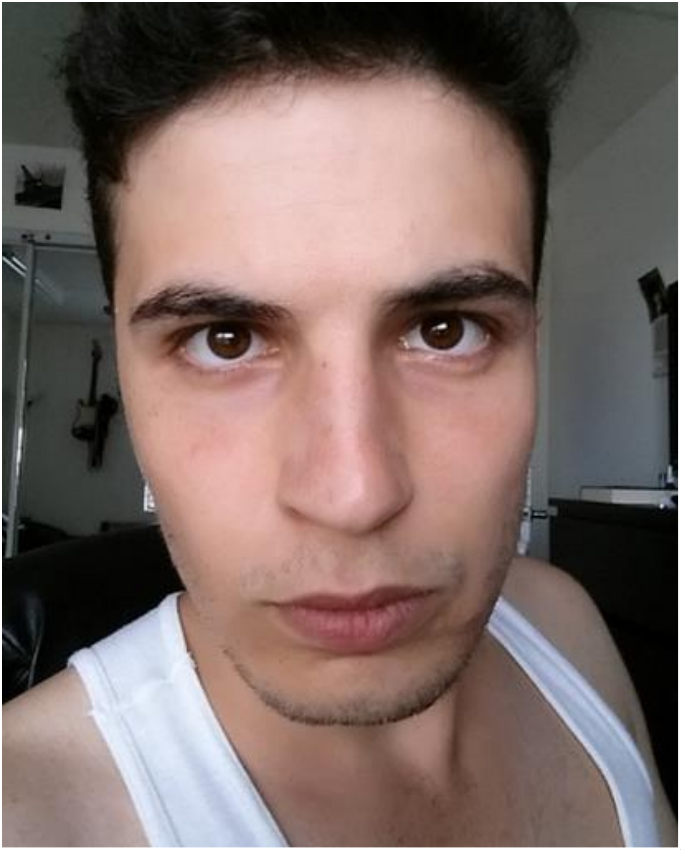


Fried et al, SIGGRAPH’16

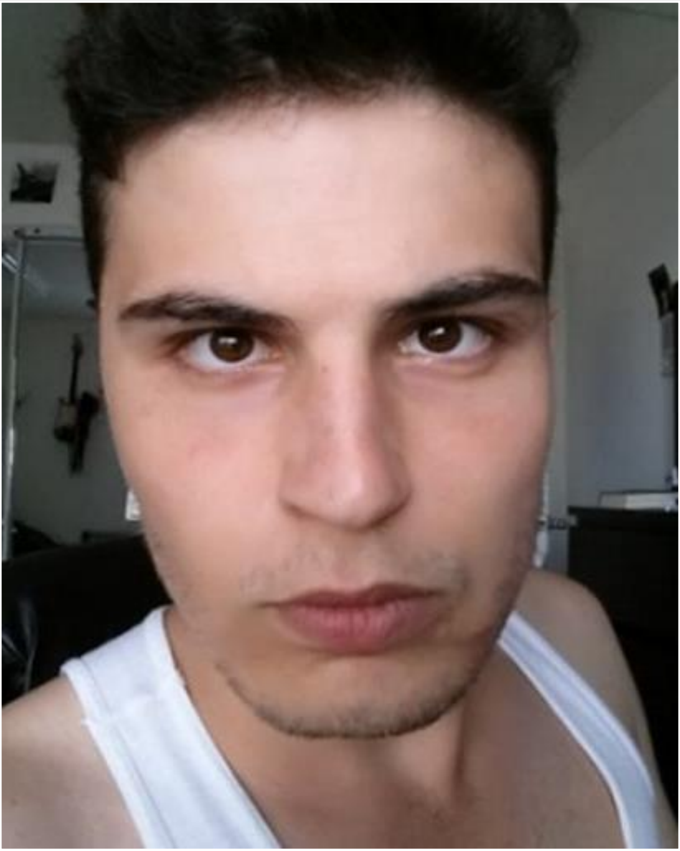


Ours

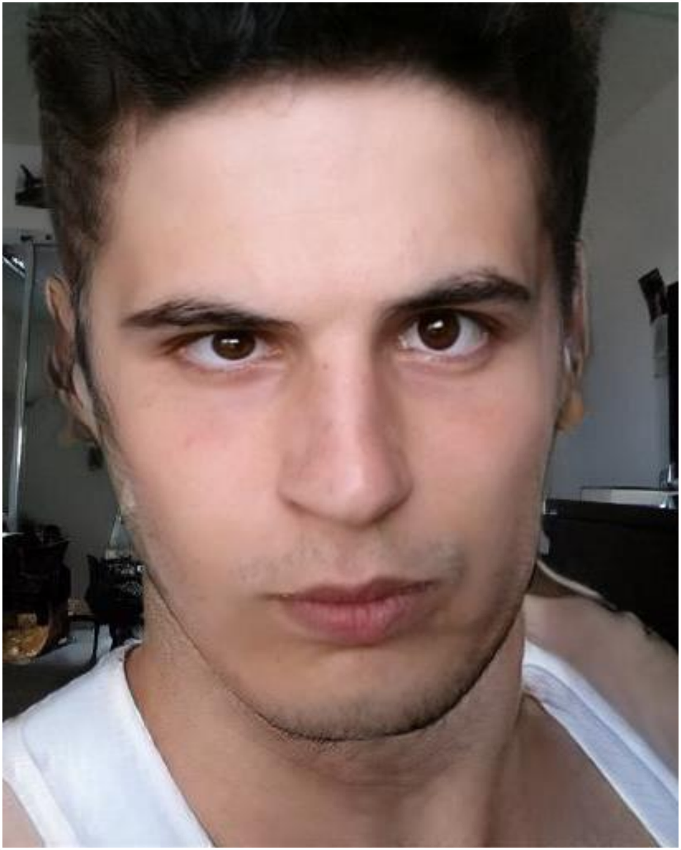
Results – Comparison



Input

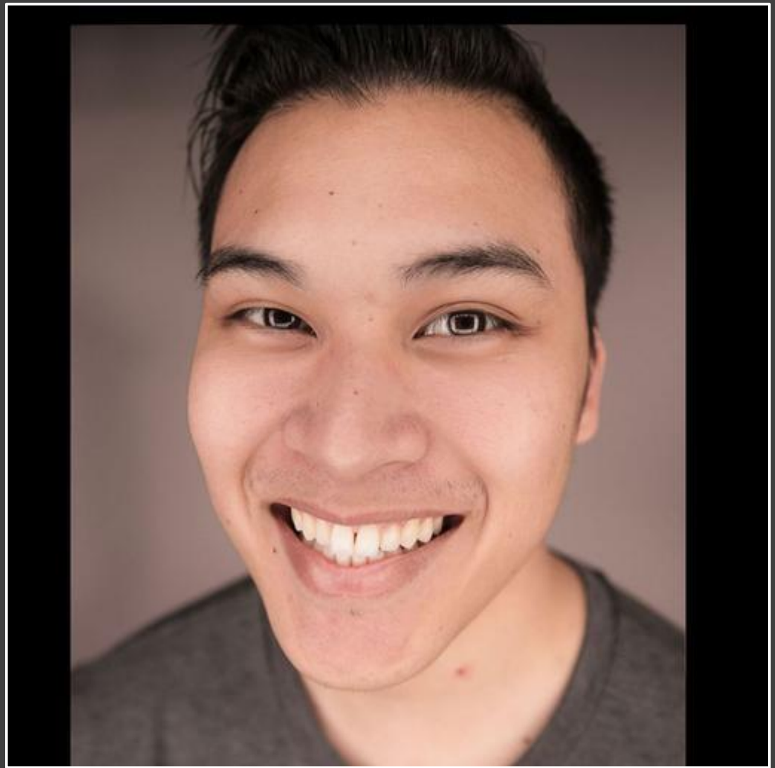


Fried et al, SIGGRAPH’16

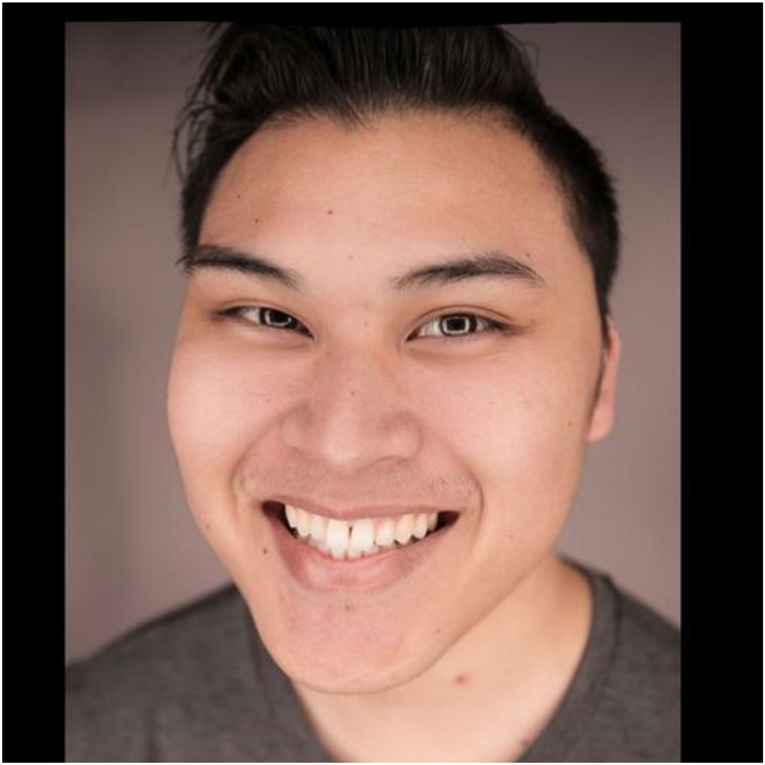


Ours

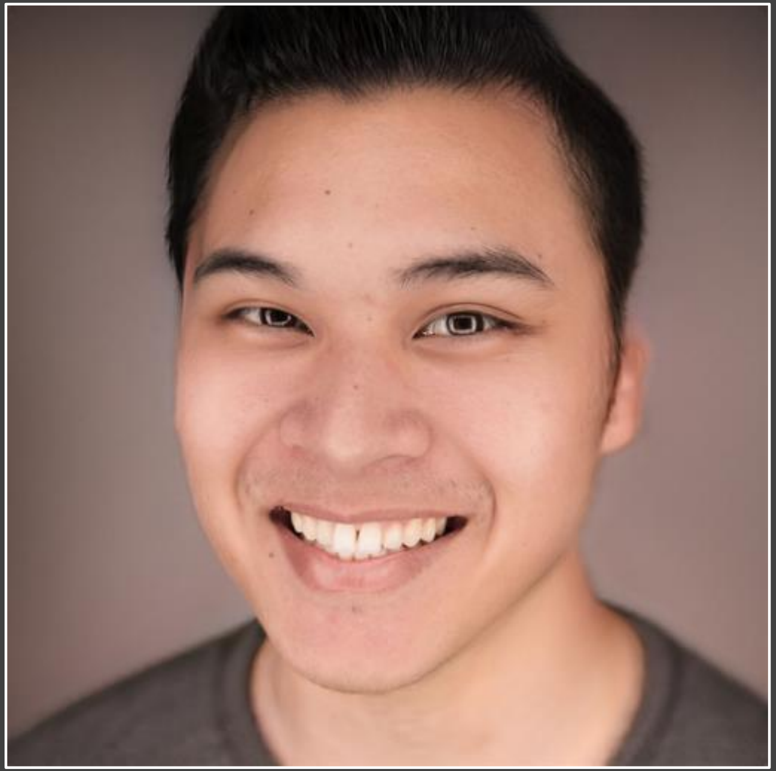
Results – Comparison



Input



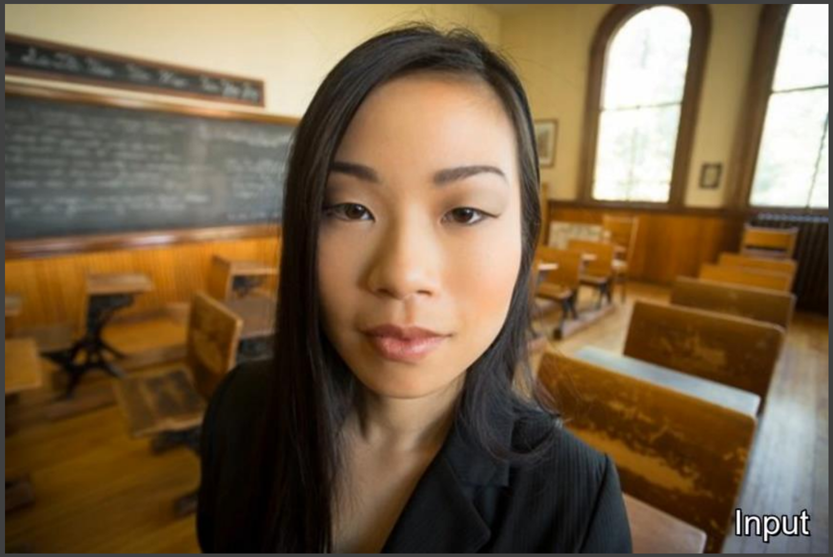
Fried et al, SIGGRAPH’16



Ours

Dolly Zoom

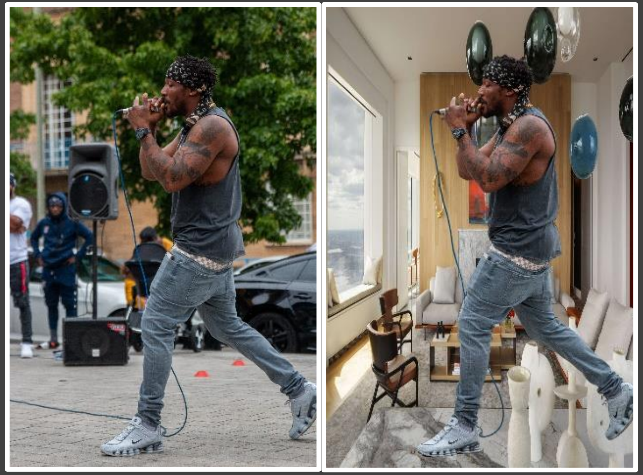
Dolly Zoom



Viewpoint + Lens

Perspective Distortion Correction

Background

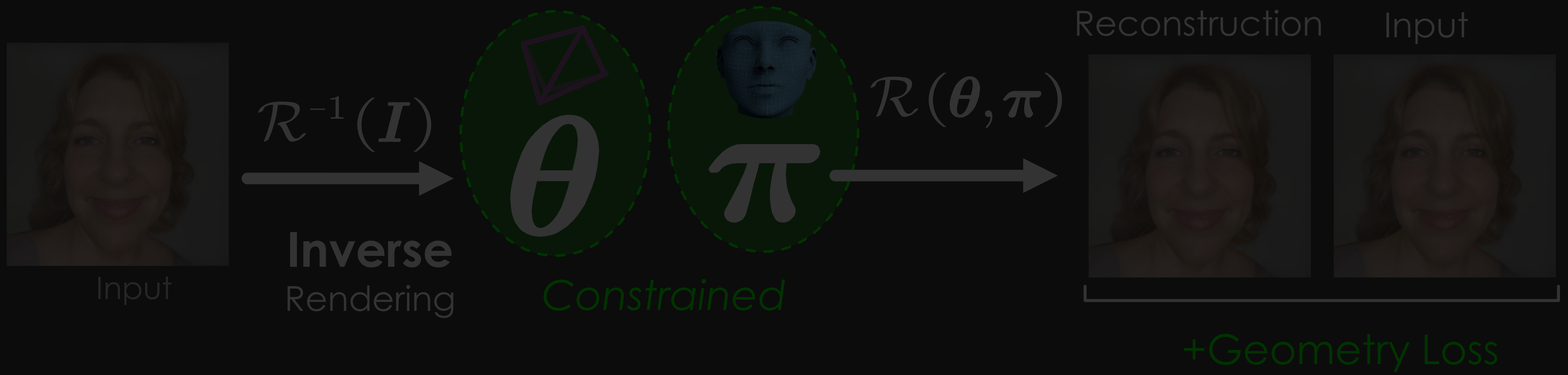


Matting by Generation

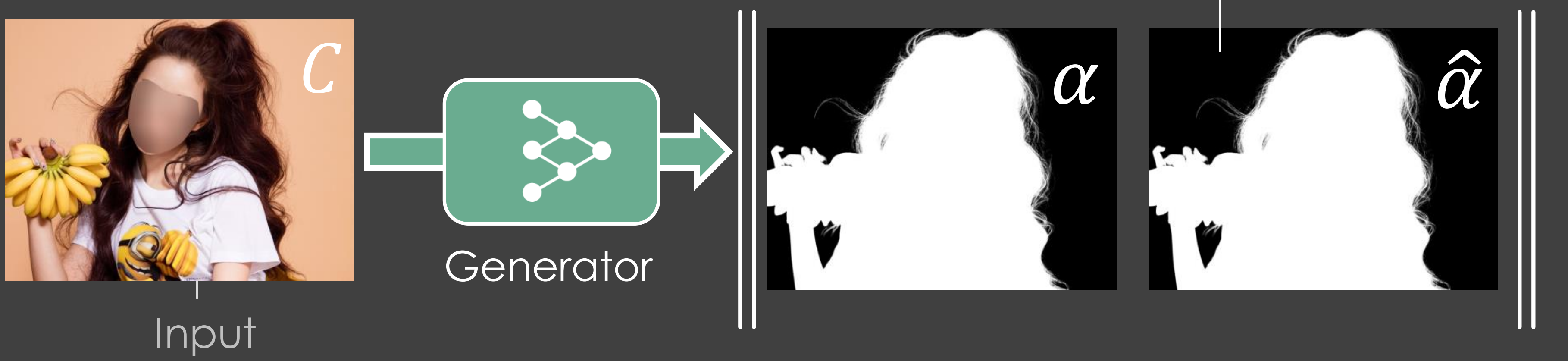
Wang et al, SIGGRAPH 2024

Harness Pre-trained Generative Models

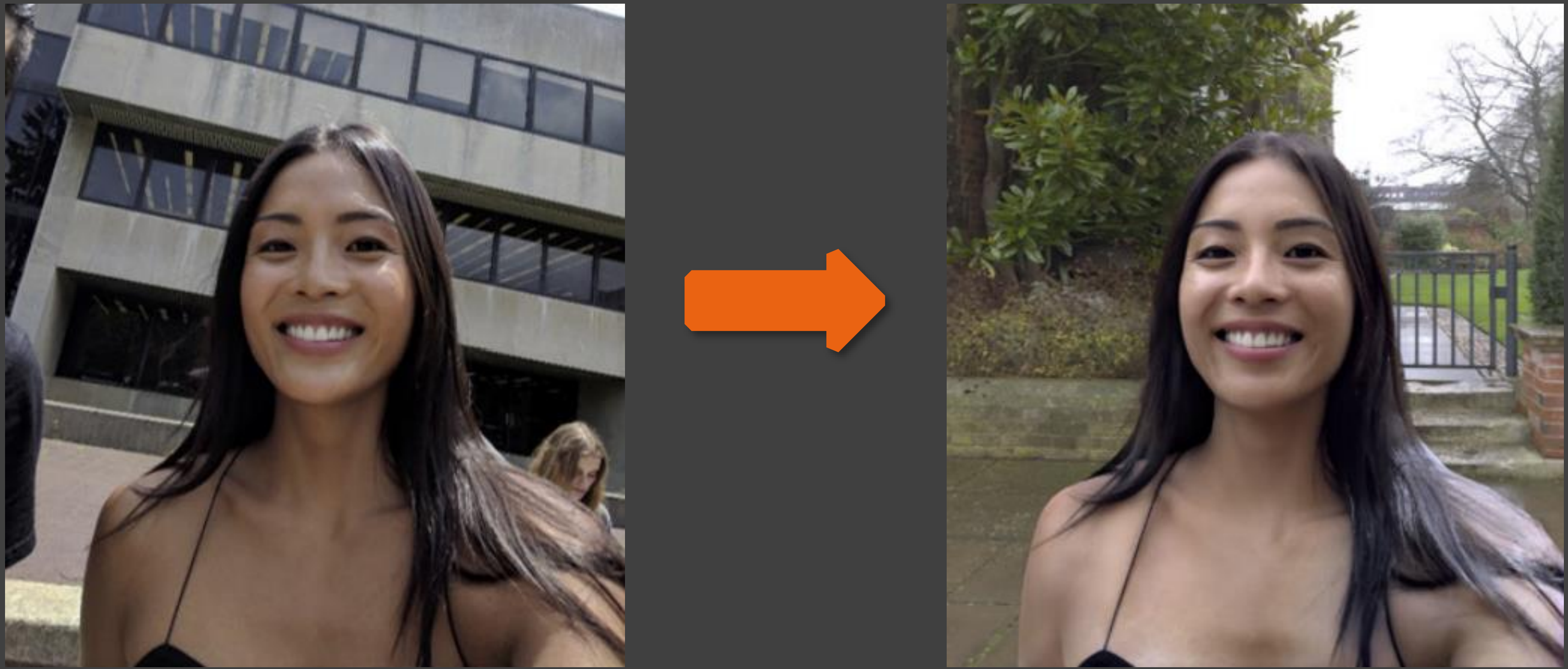
Optimization-based: no labels required



Learning with Labels: imperfect labels



Manipulate Background



Background Gallery

Image credit: total relighting

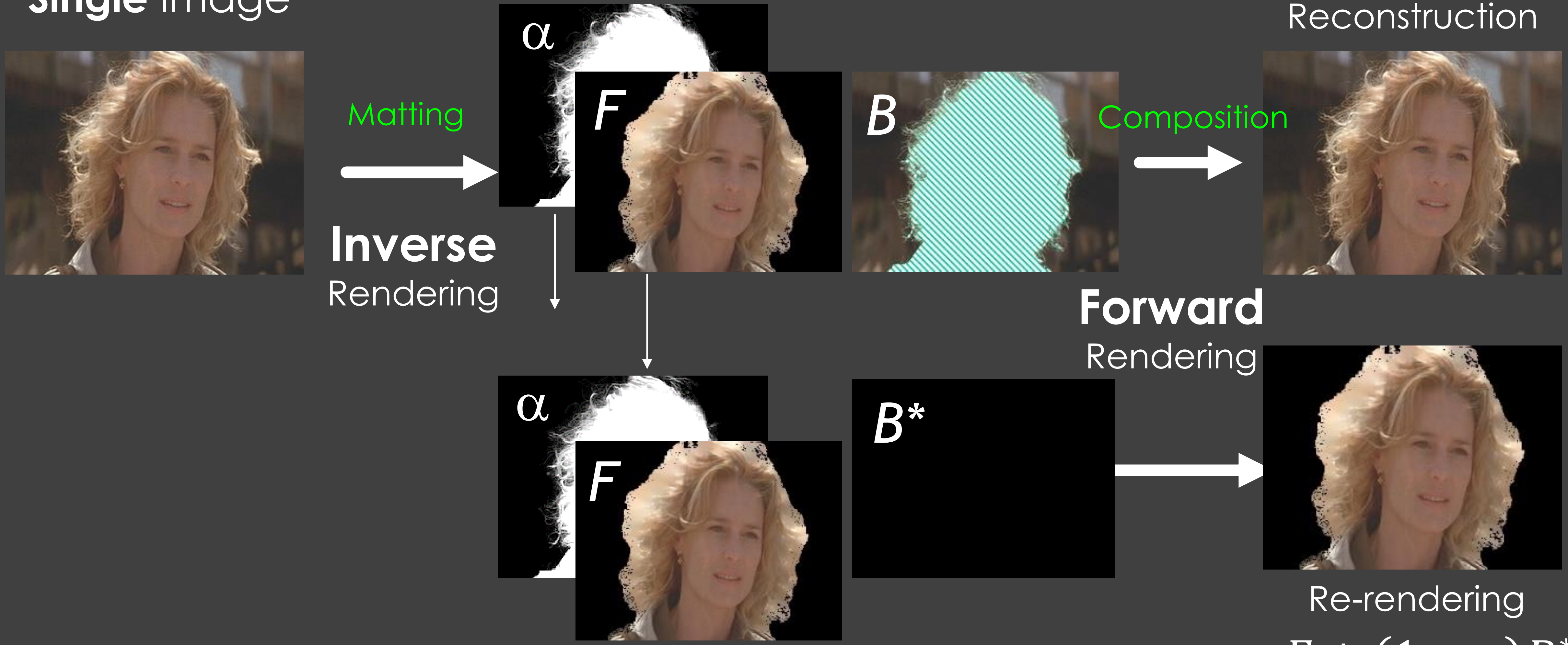
Factorization Problem

Input: I

Single Image

$\alpha F + (1-\alpha)B$

Reconstruction



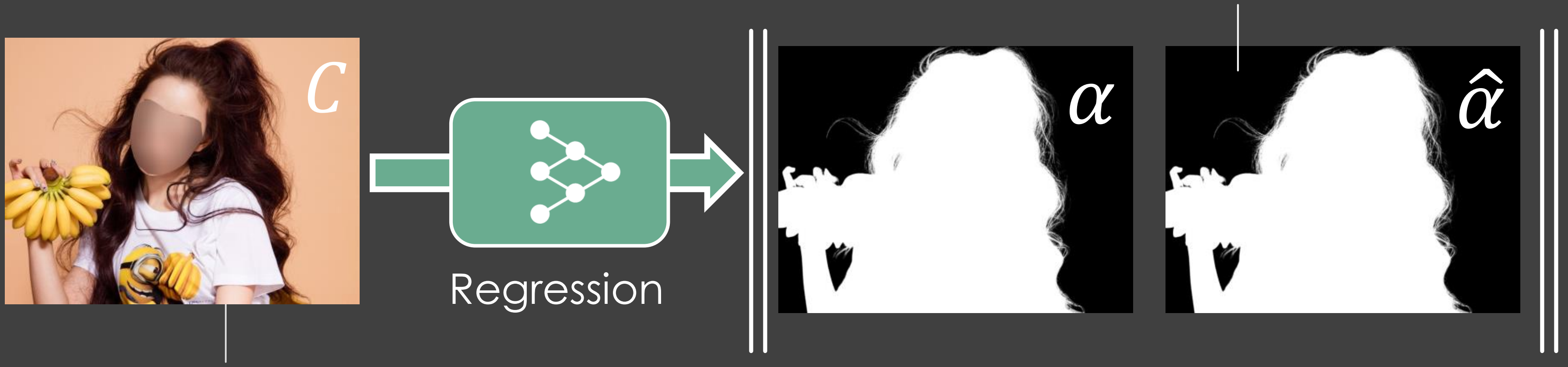
Inverse Rendering

Forward Rendering

Re-rendering

$\alpha F + (1-\alpha)B^{*}$

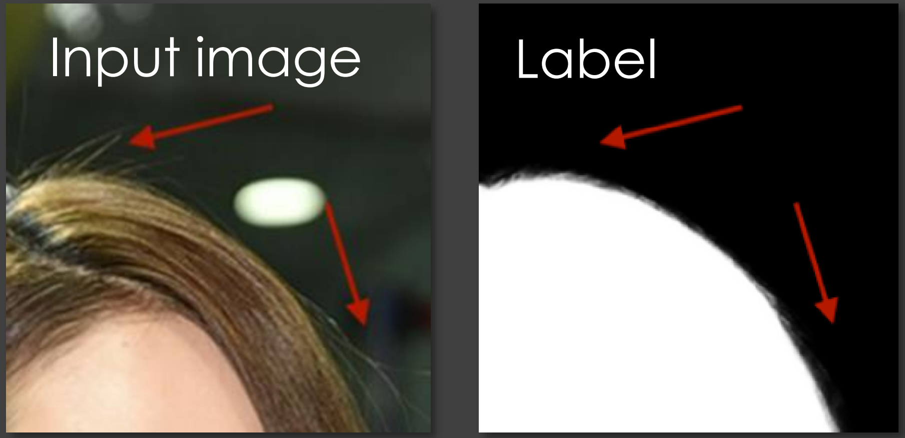
Learning with Labels



Human Annotations

Regression

Input

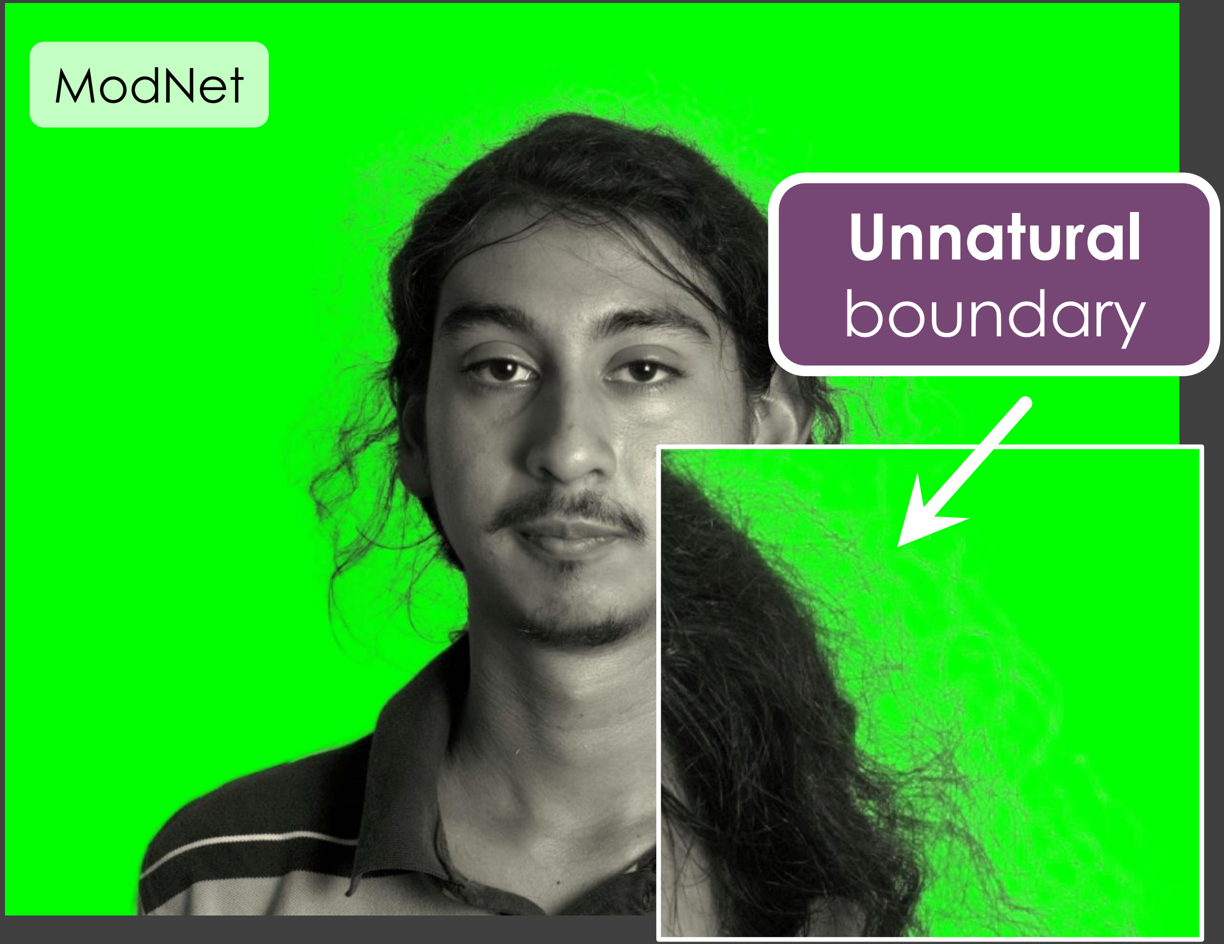

Input image

Label

Poor label quality

Ke et al, MODNet, AAAI'22
Li et al, P3M, MM'22
Ma et al, ViTAE-S, IJCV'23

Limitations of Existing Methods



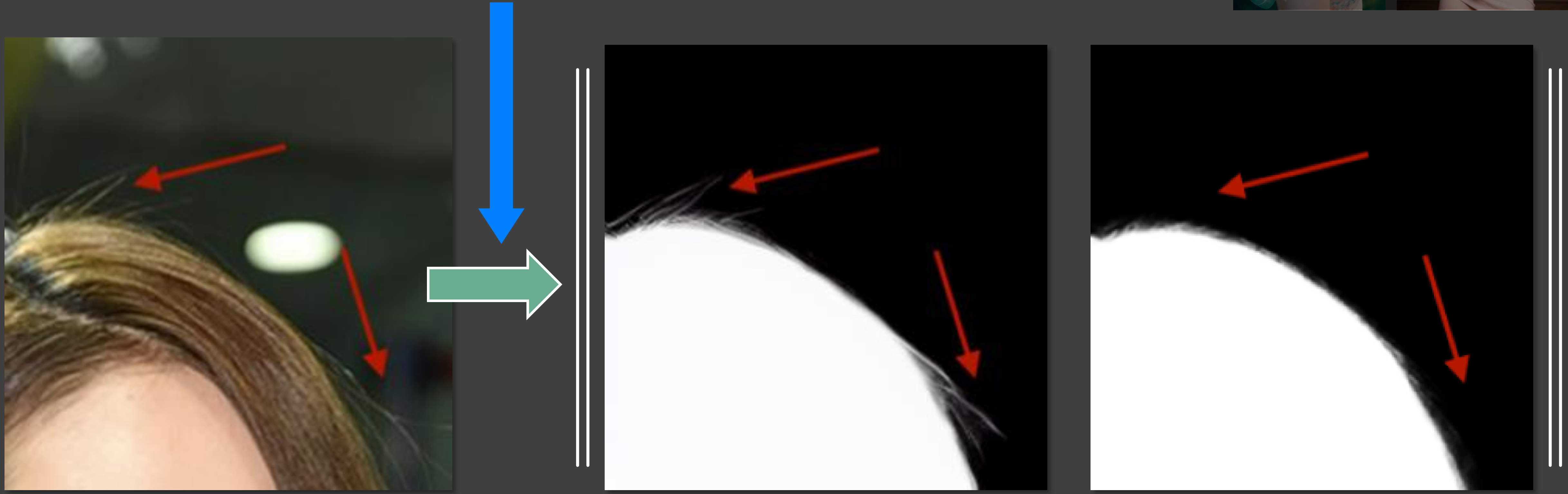
Unnatural boundary

Ke et al, MODNet, AAAI’22

Generative Diffusion Prior

Diffusion Model with Rich Image Prior

Generative Prior for Regularization

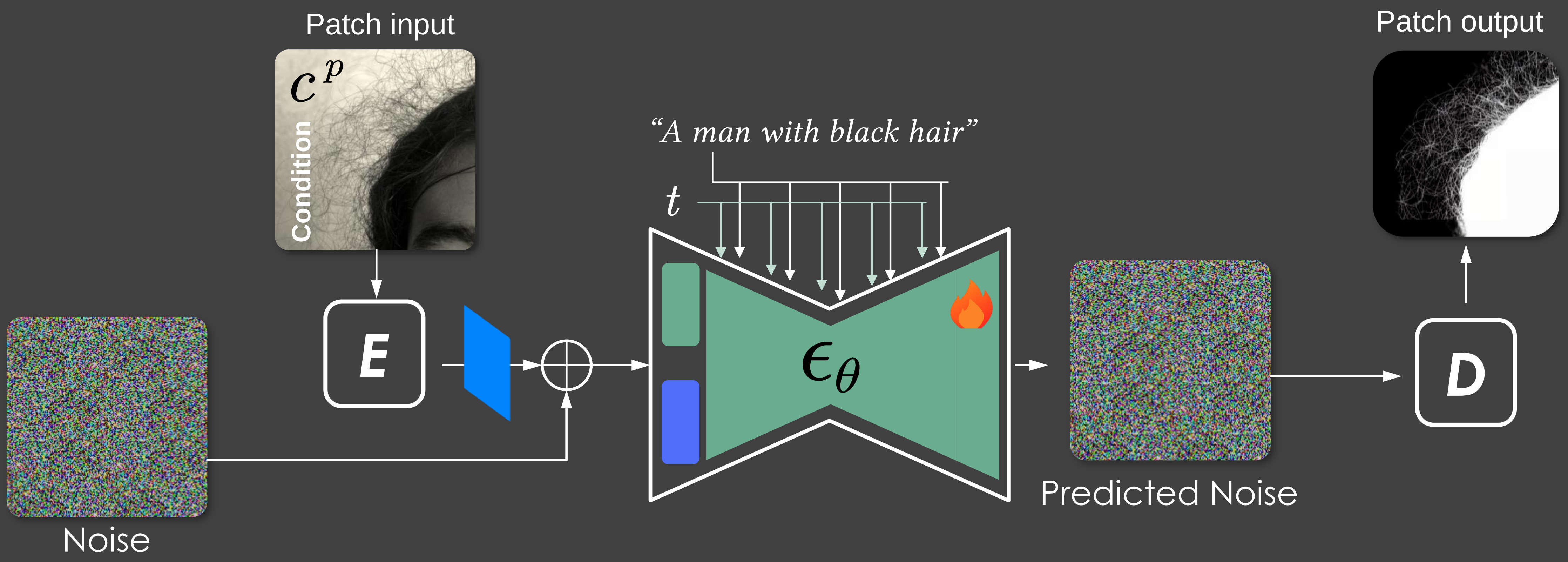


Input training image

Output

Label

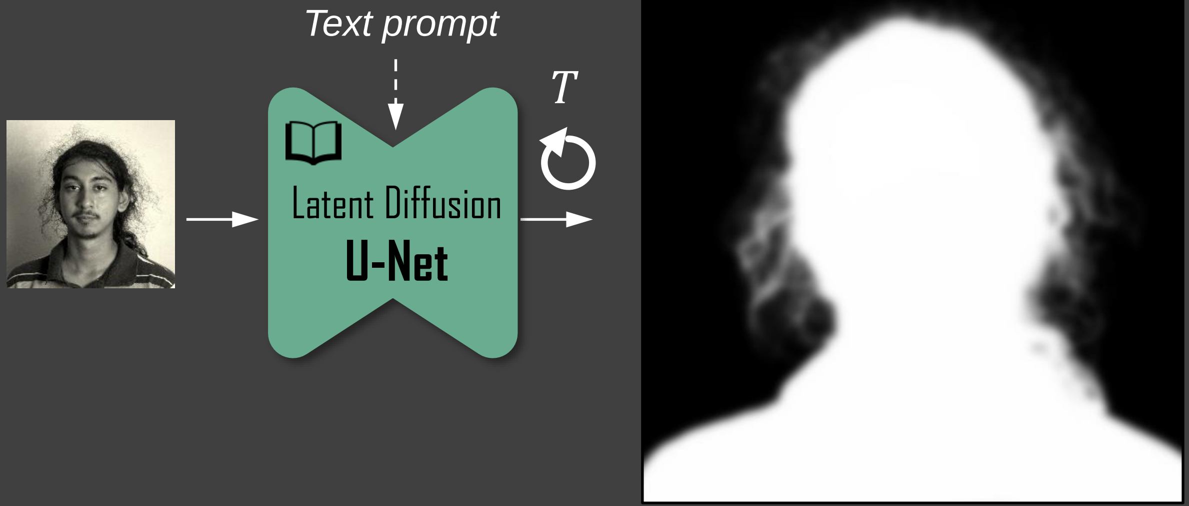
Repurposing Latent Diffusion Model



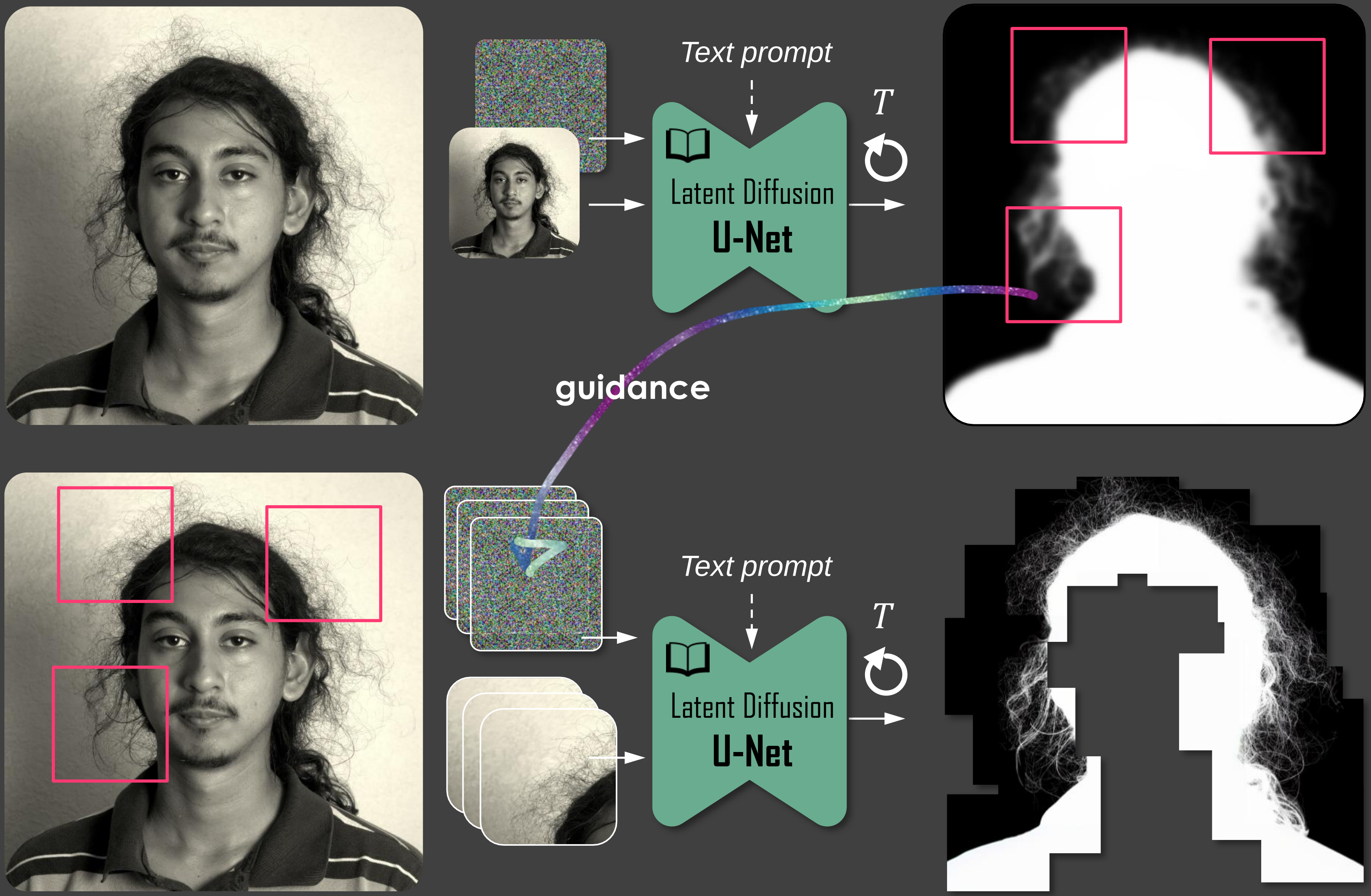
🟩 Original weight

🟦 New weight

Challenge of Processing HR Images

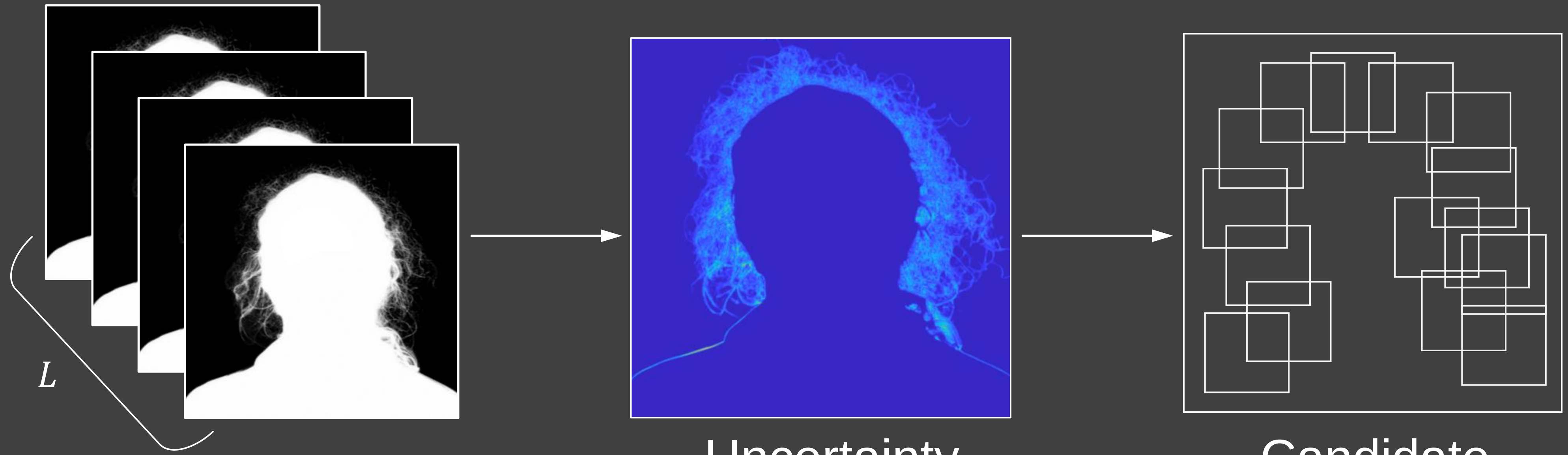


Pipeline for Processing HR Images



Pipeline for Processing HR Images

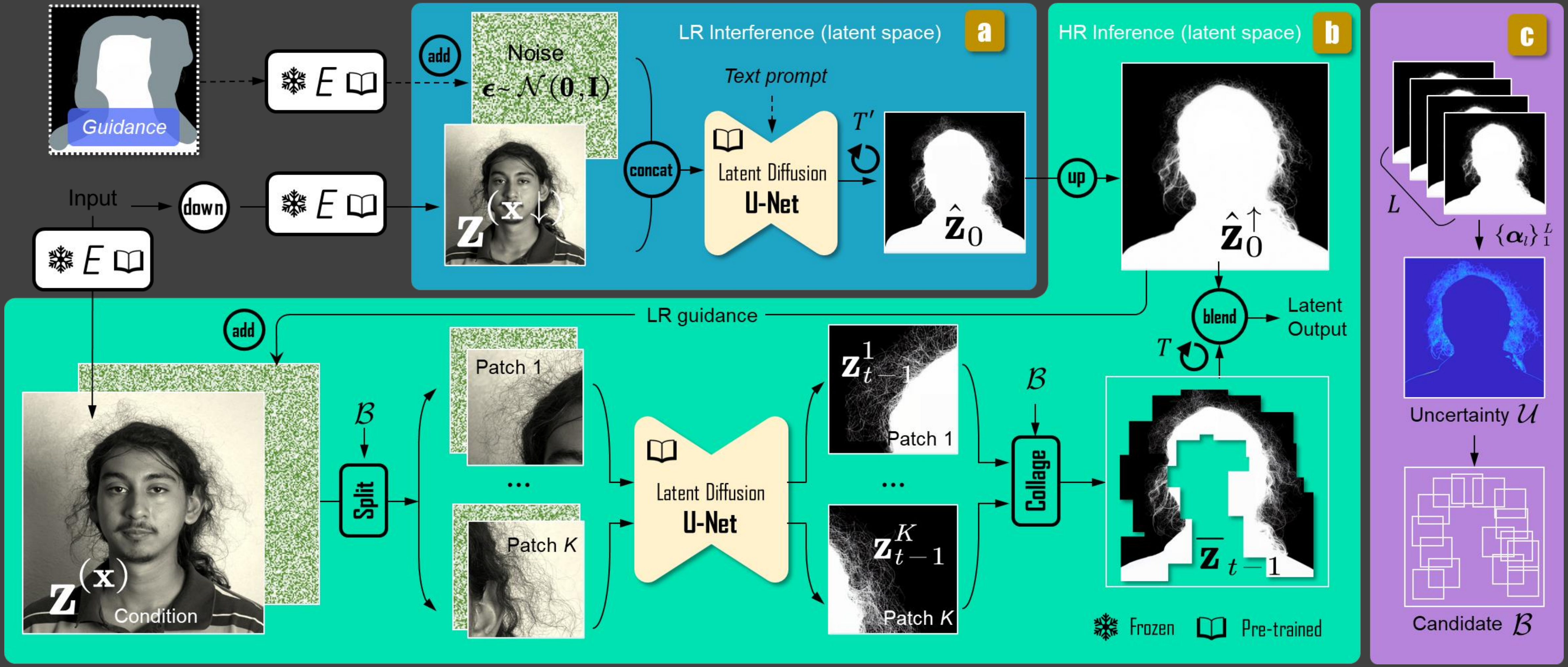
Get potential areas by uncertainty

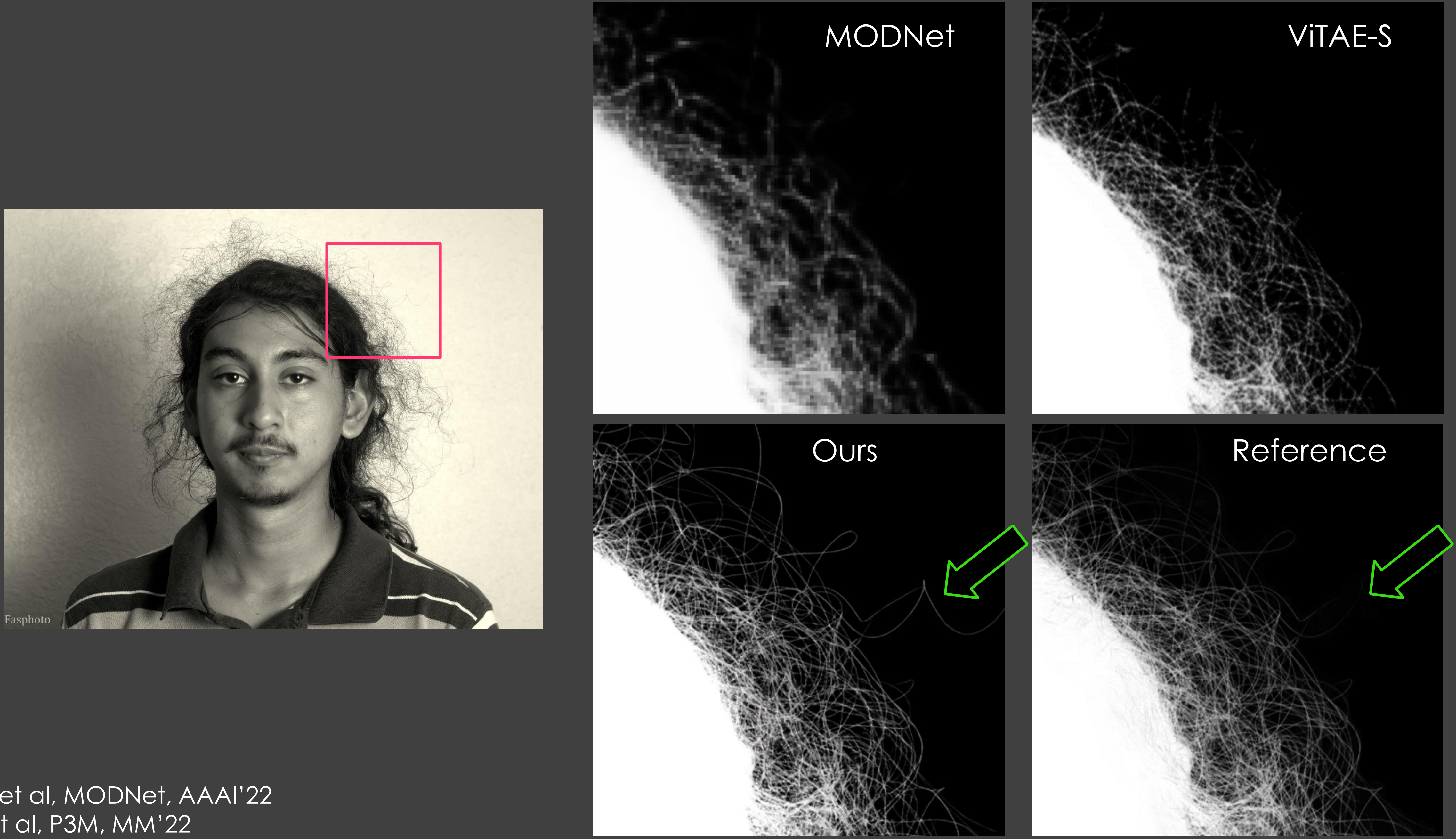


Uncertainty

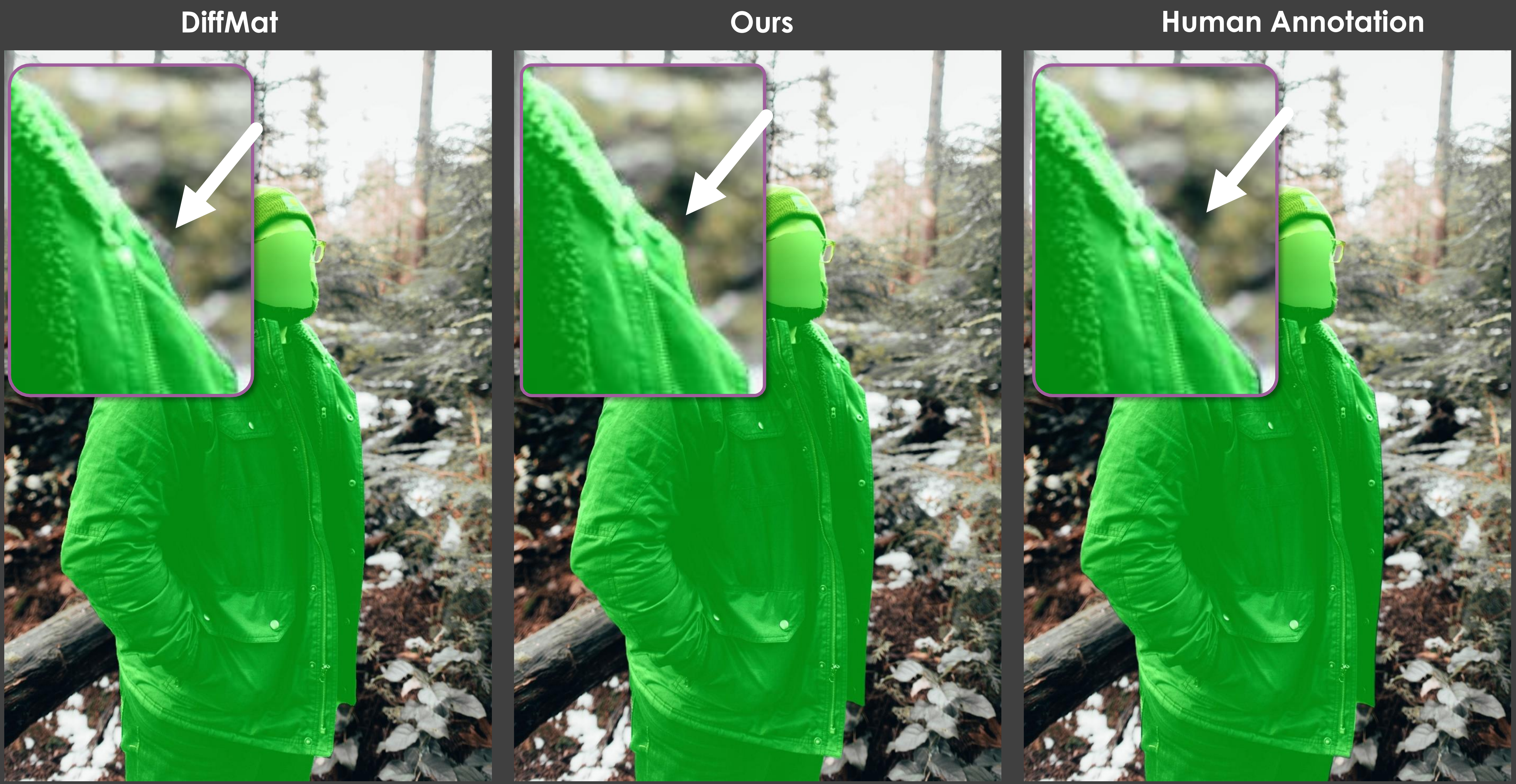
Candidate

Full Pipeline





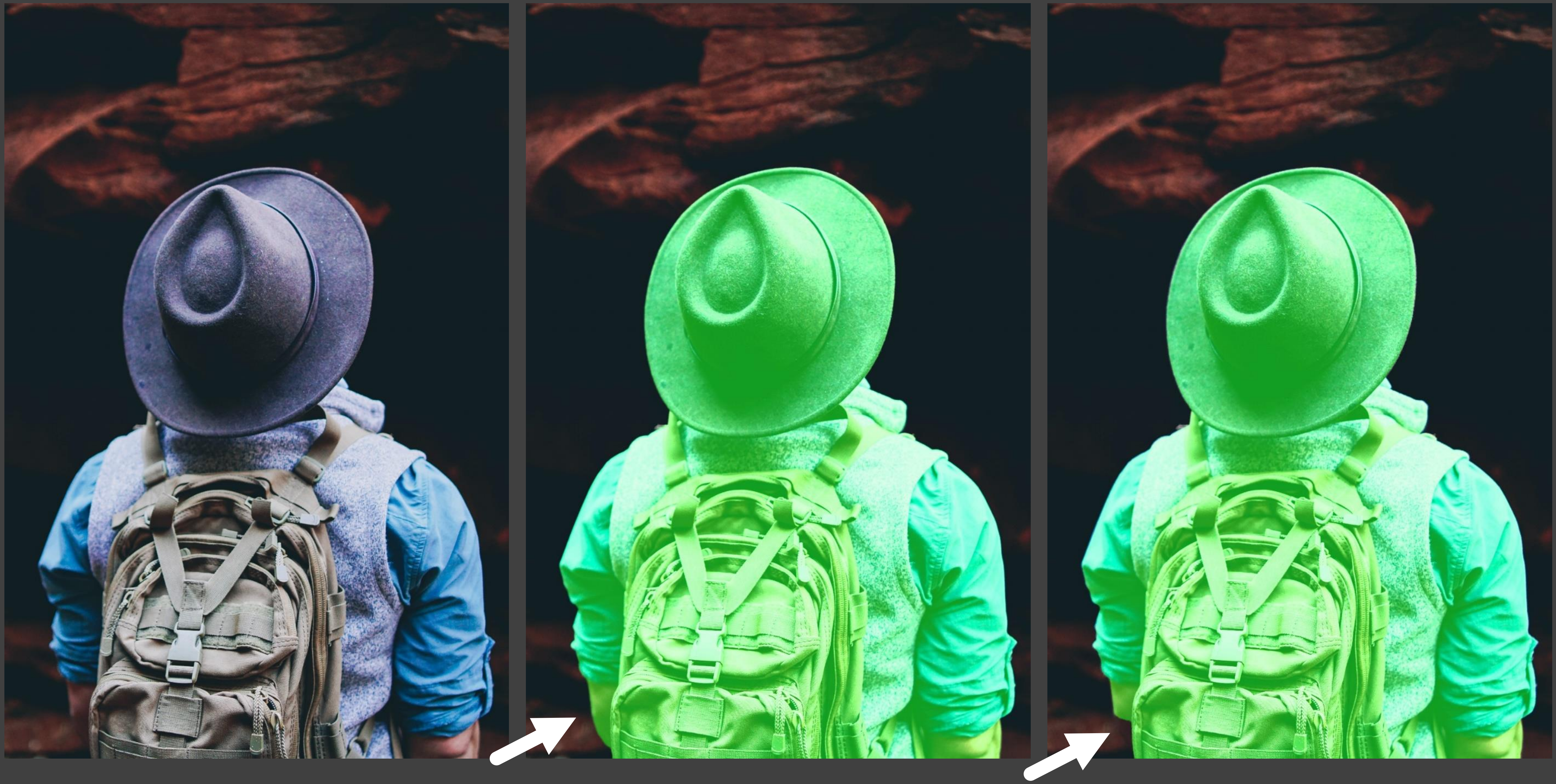
Ke et al, MODNet, AAAI’22
Li et al, P3M, MM’22
Ma et al, ViTAE-S, IJCV’23


DiffMat
Ours
Human Annotation

Input

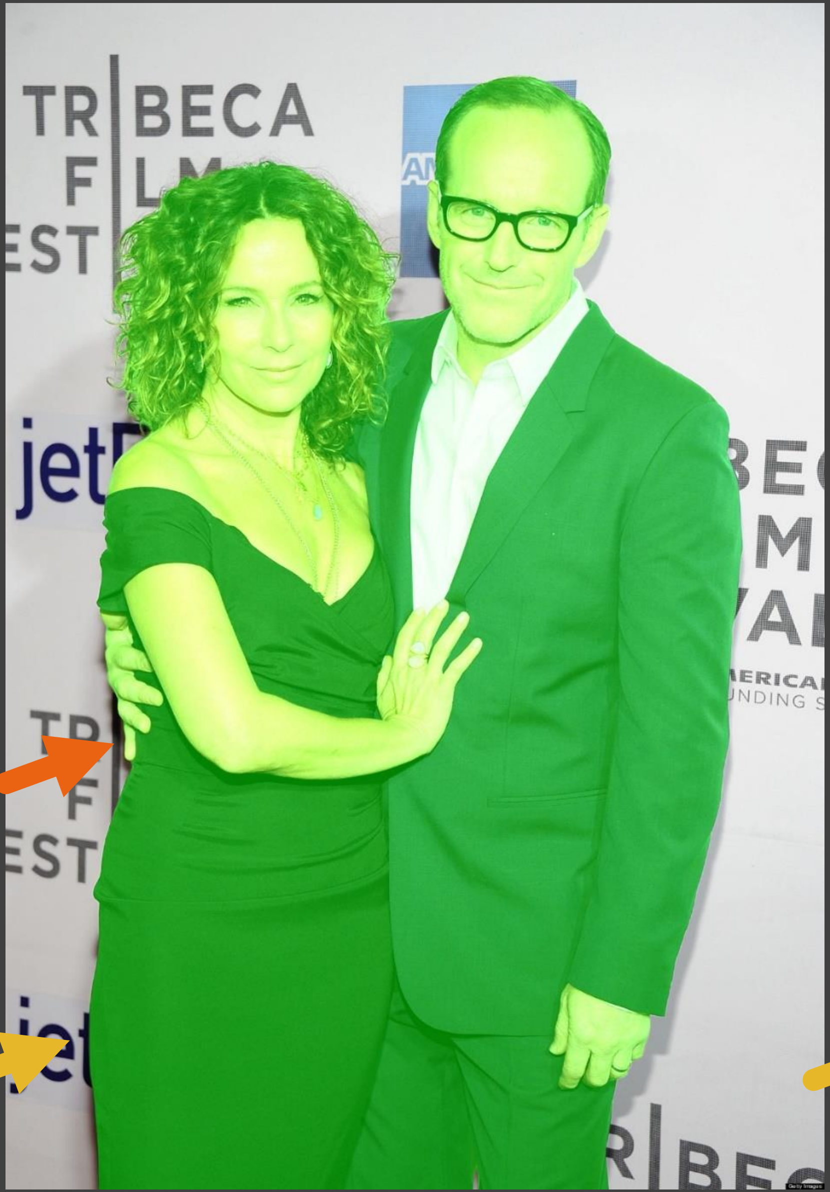
Ours

Human Annotation

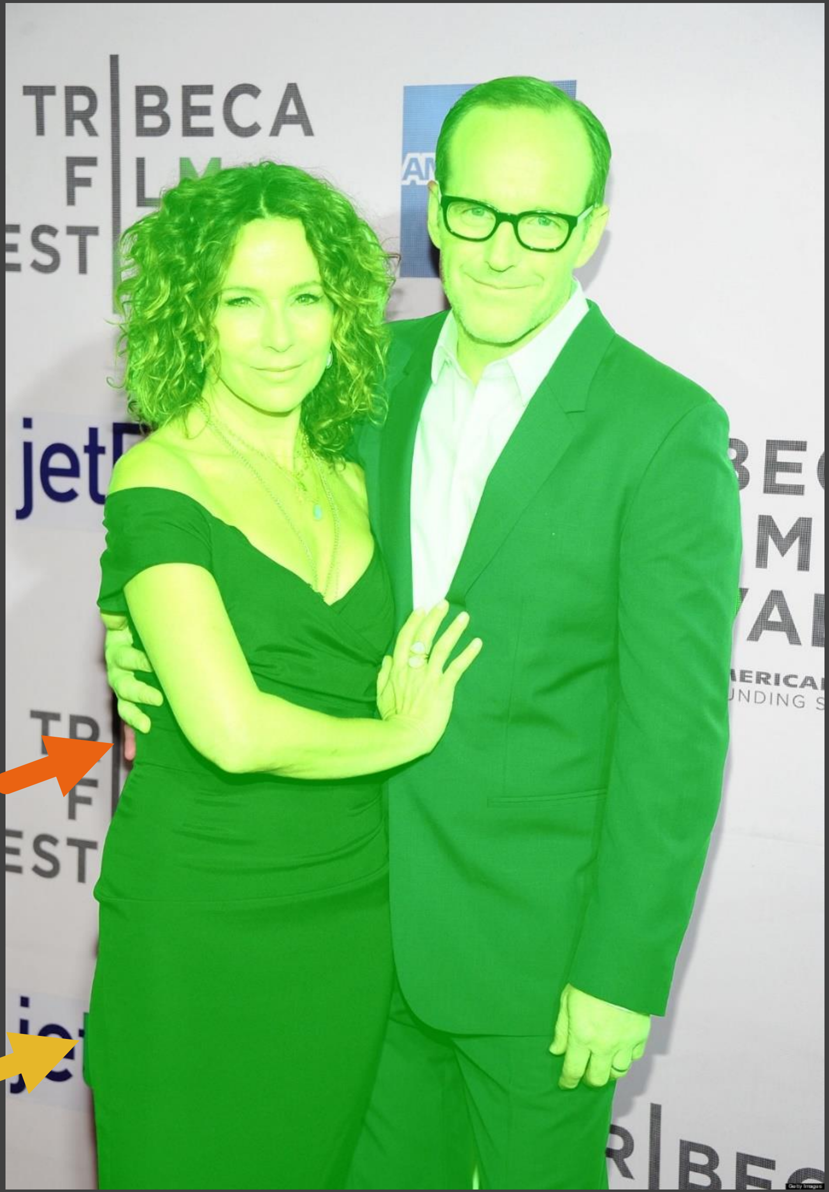


Input

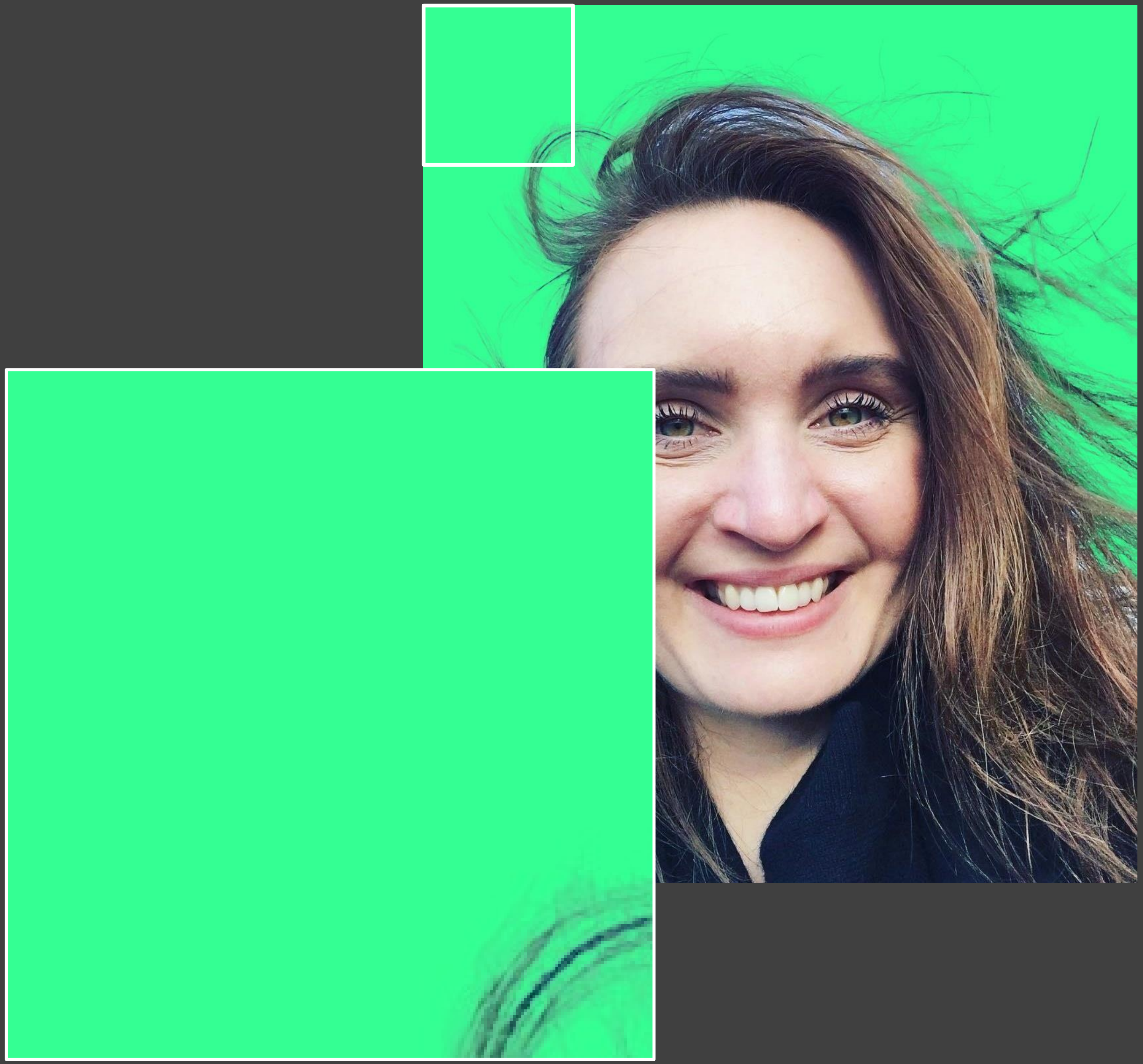
Ours



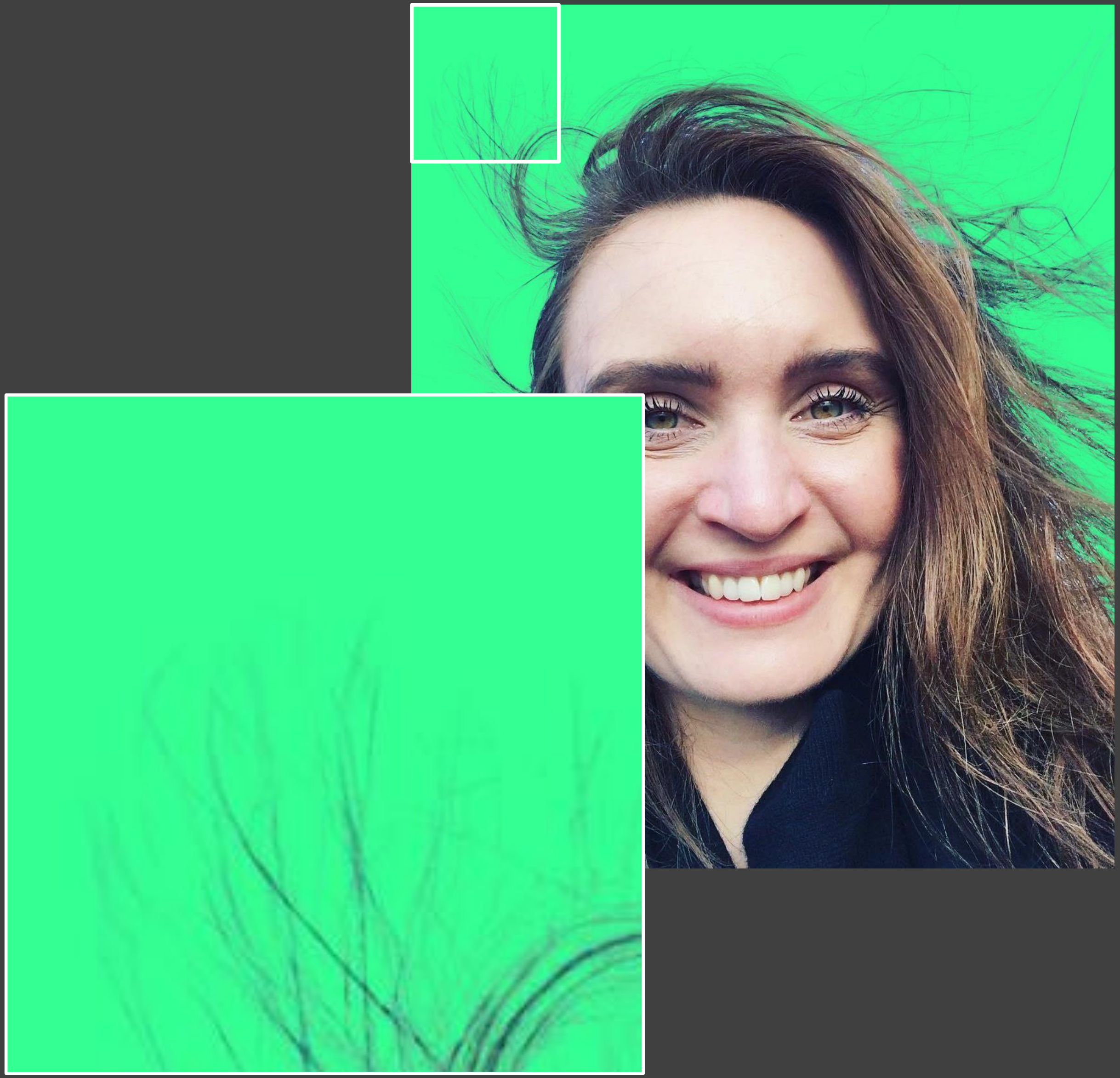
Human Annotation

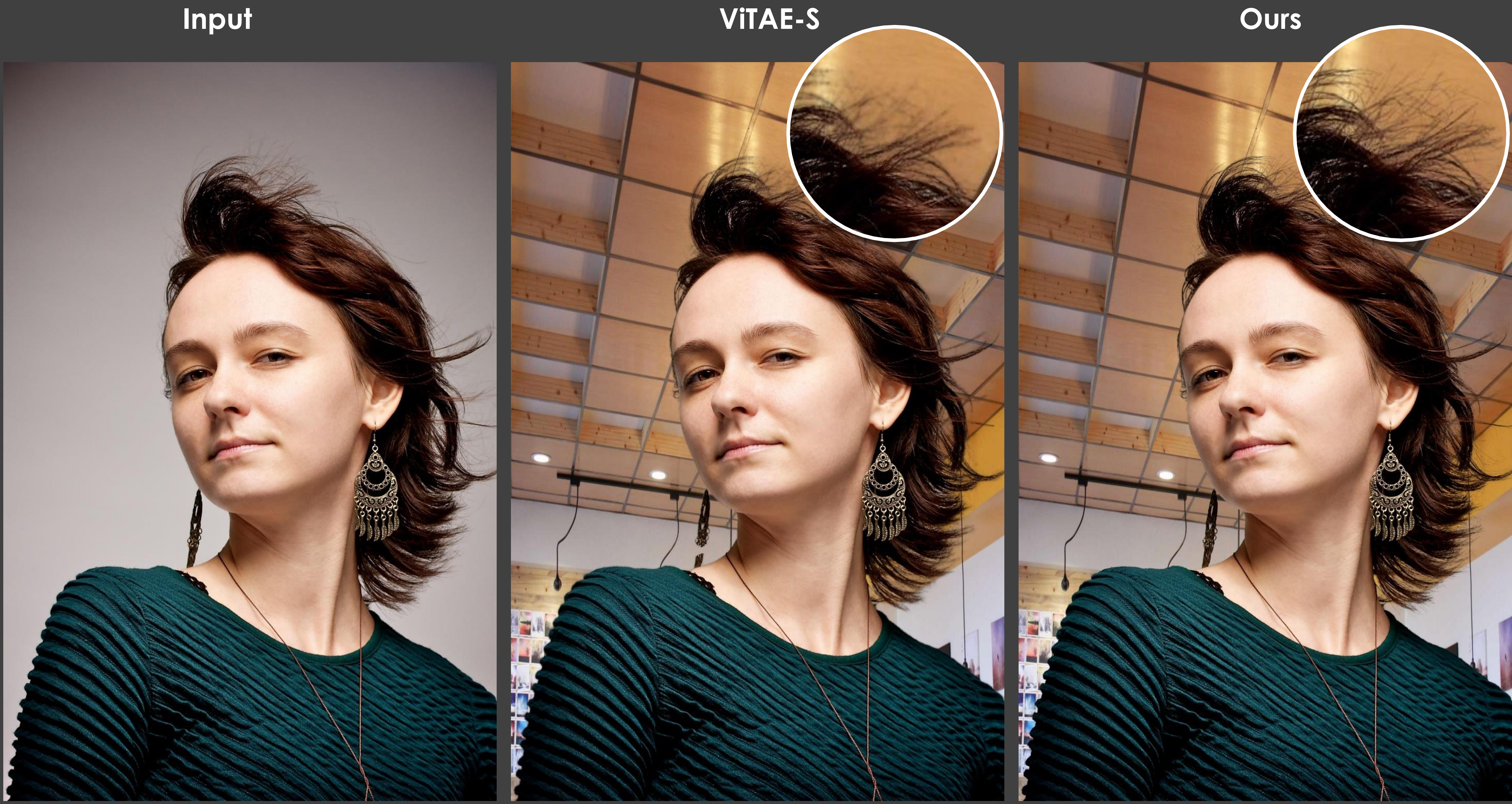


ViTAE-S



Ours

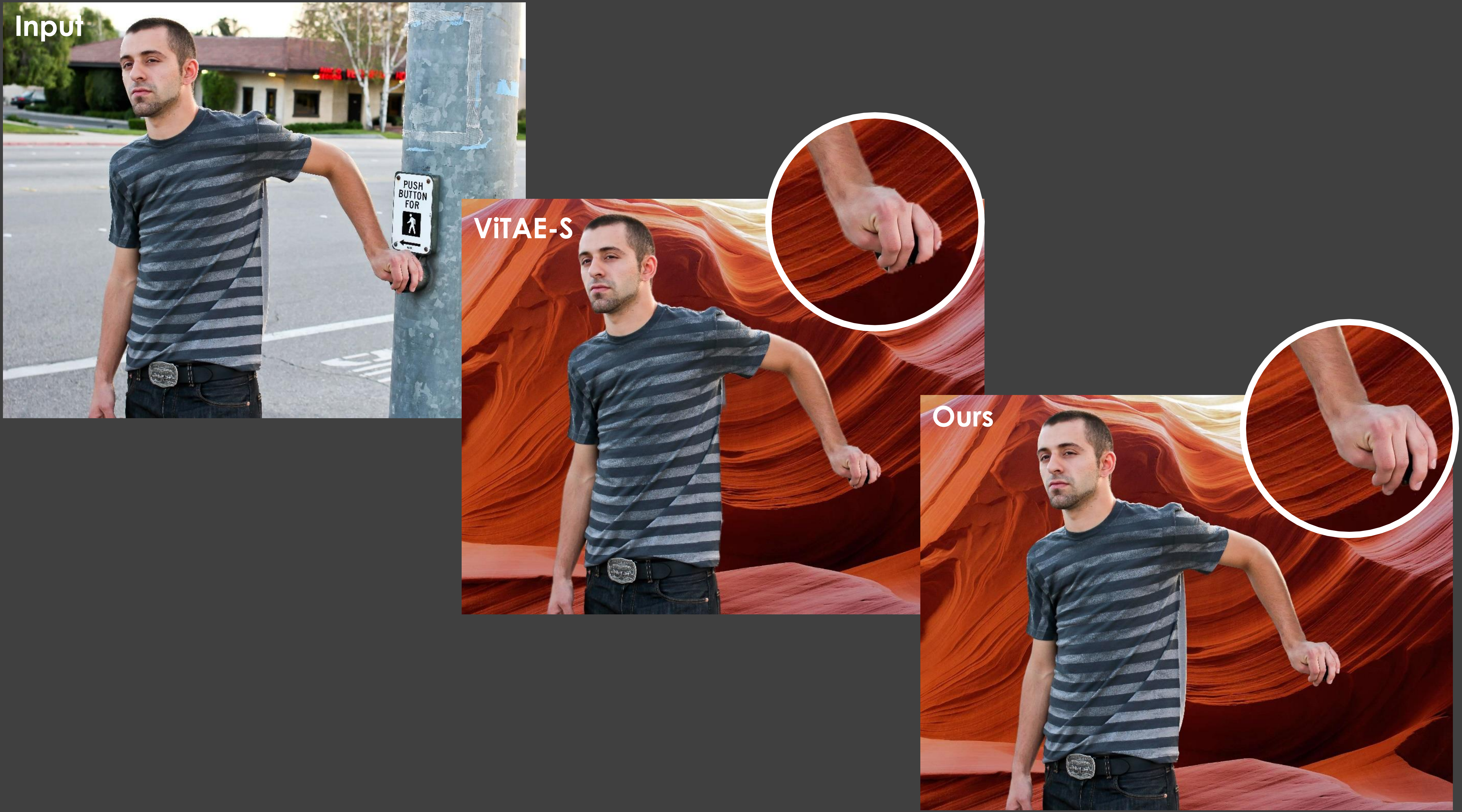


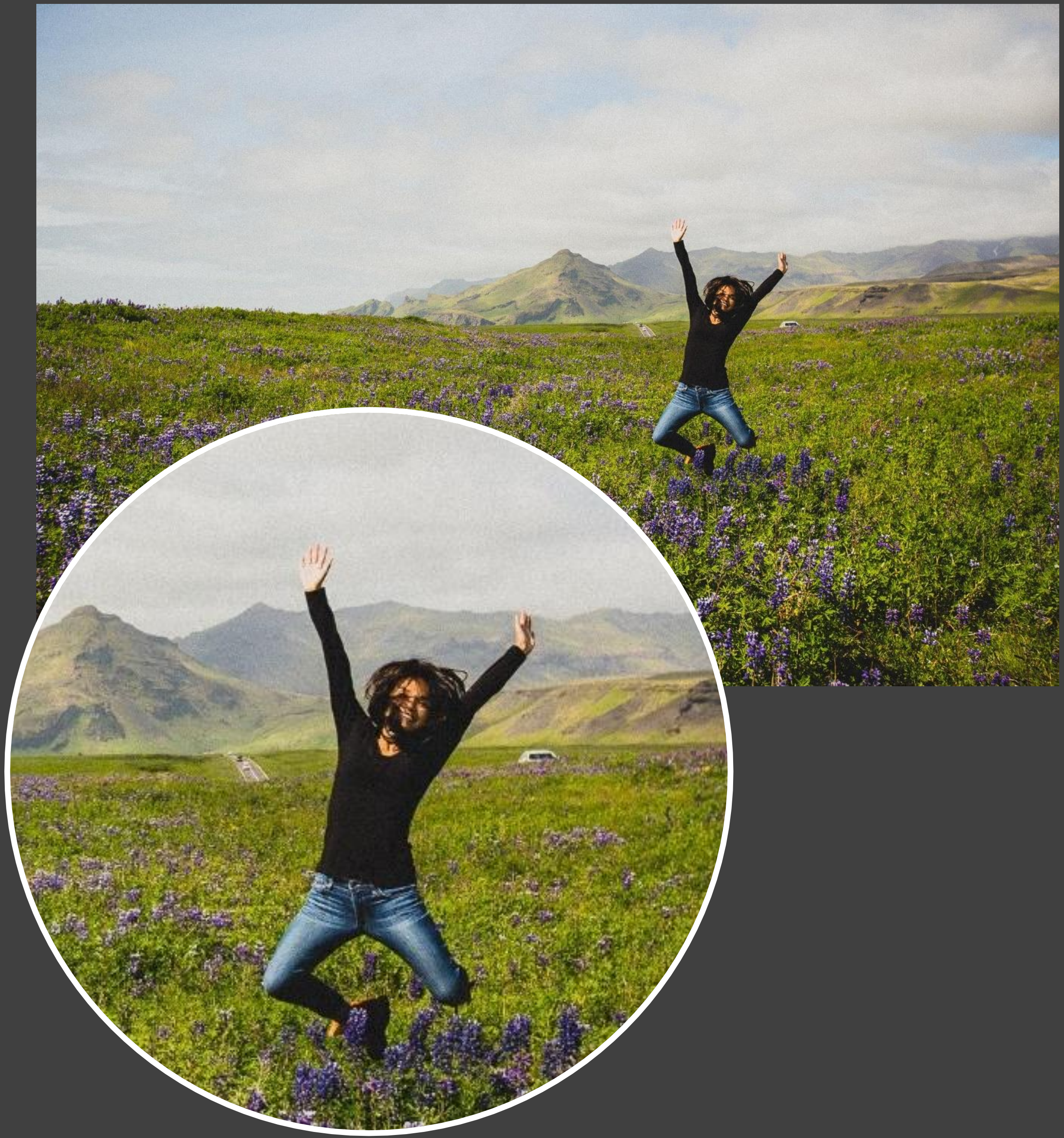


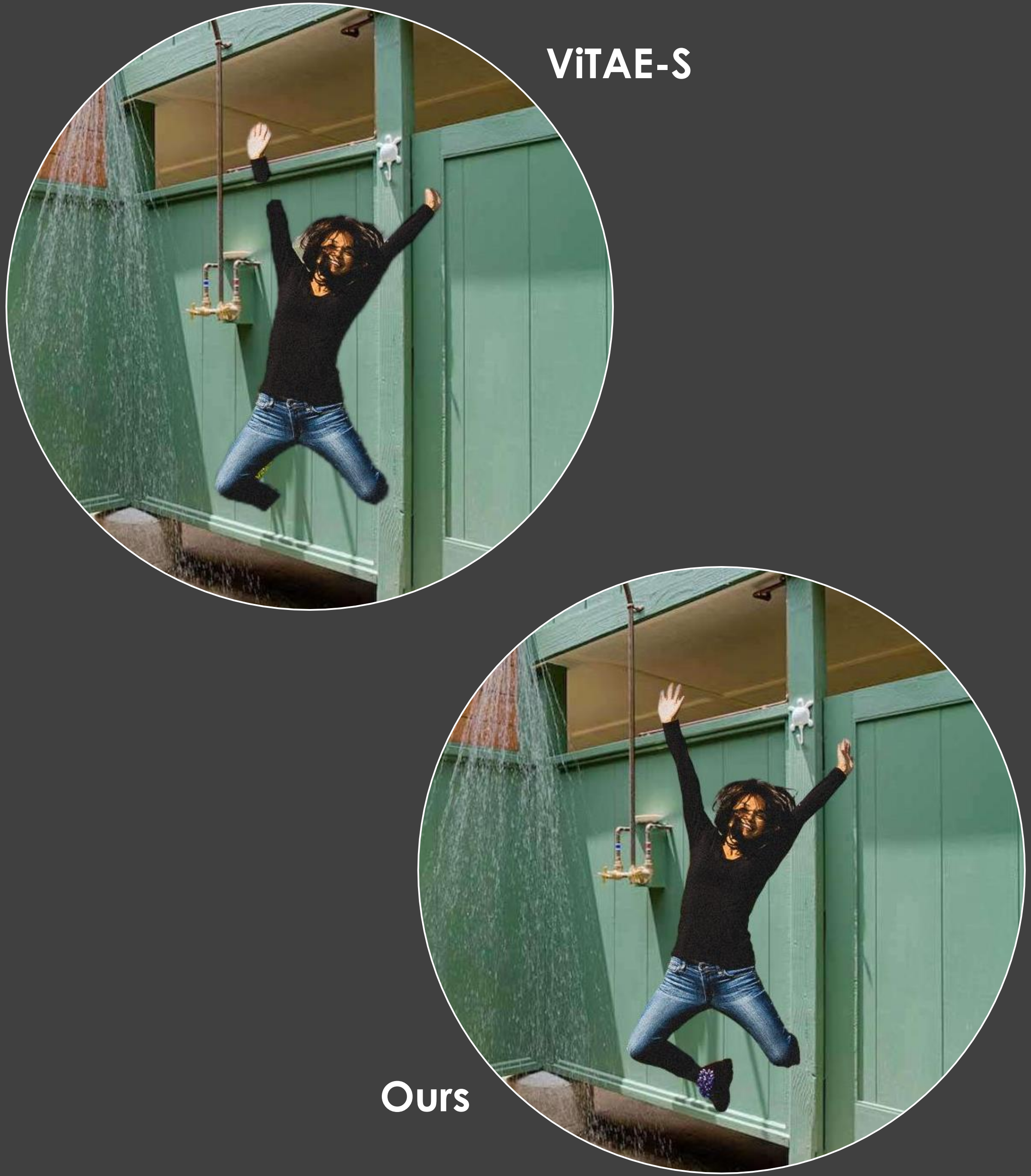
Input

ViTAE-S

Ours


Input
PUSH BUTTON FOR
ViTAE-S
Ours

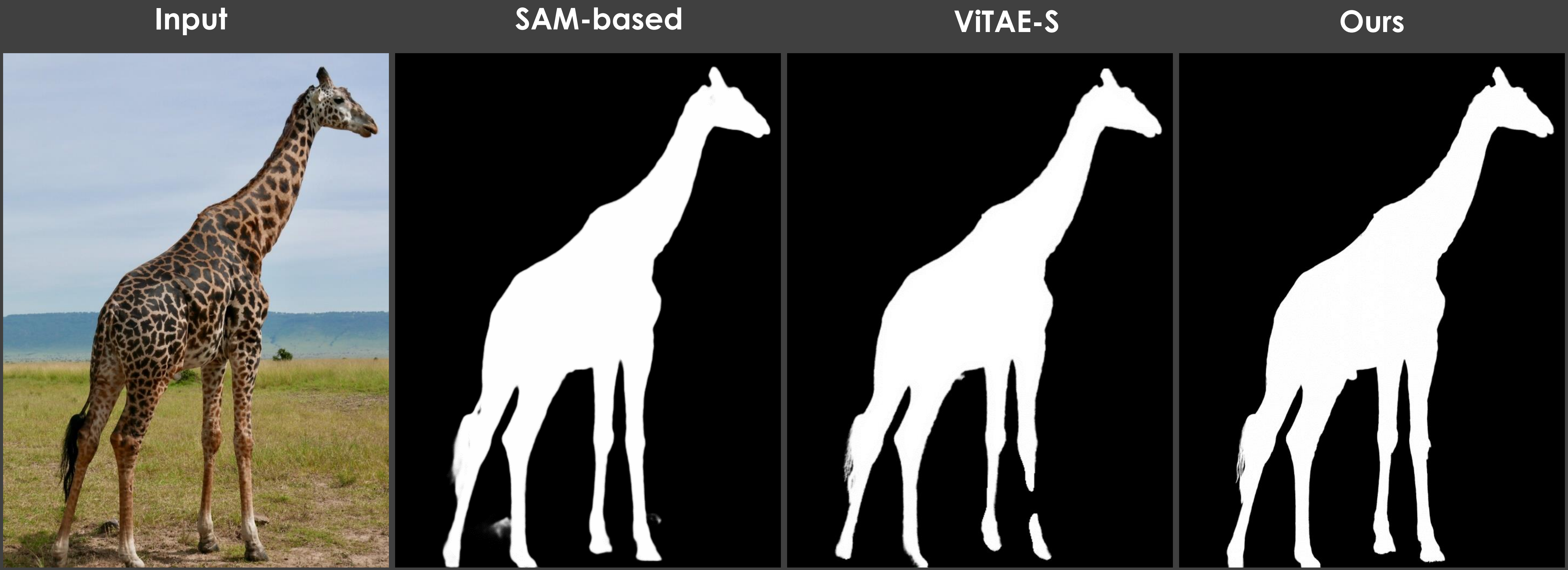




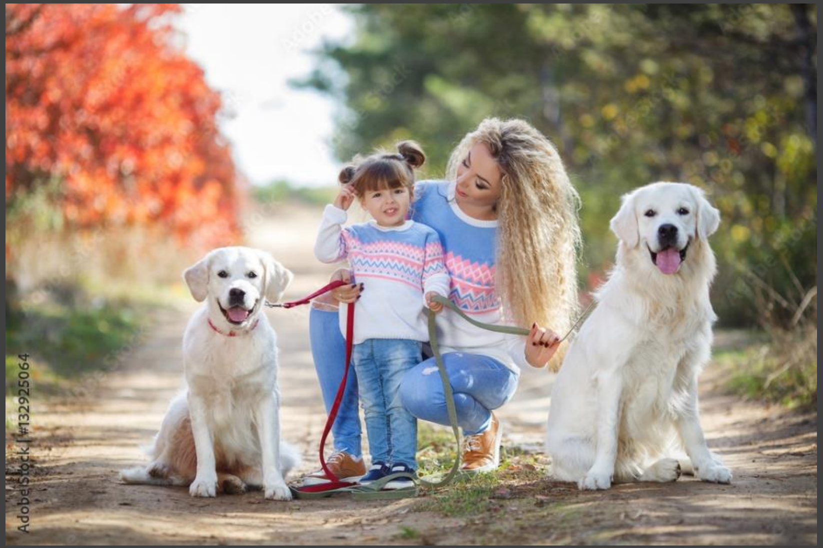
ViTAE-S

Ours

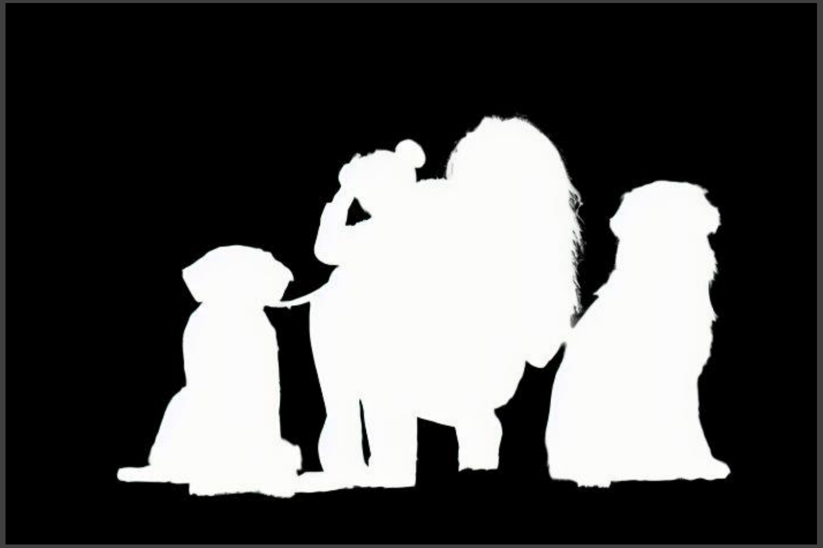
Out-of-Distribution Matting



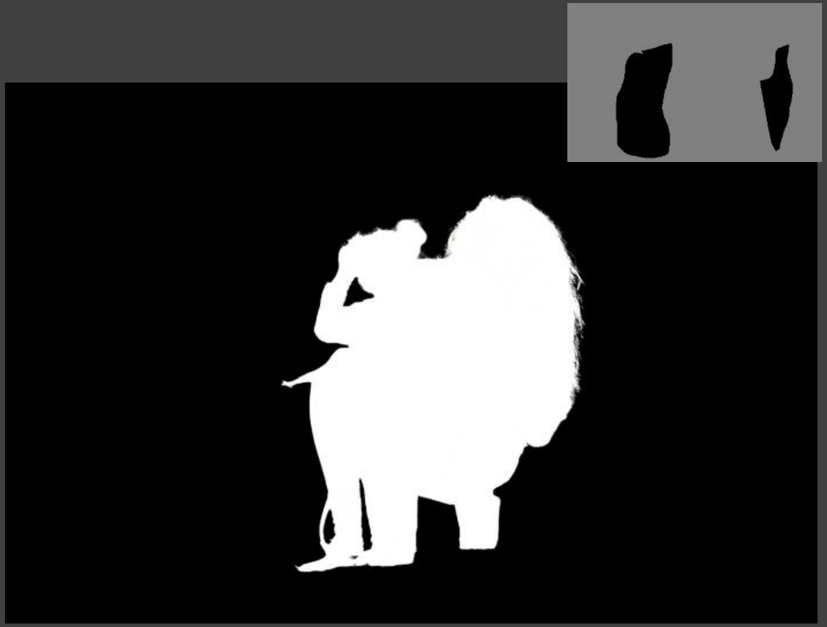
Matting with Additional Guidance



Input



w/o guidance



w/ guidance

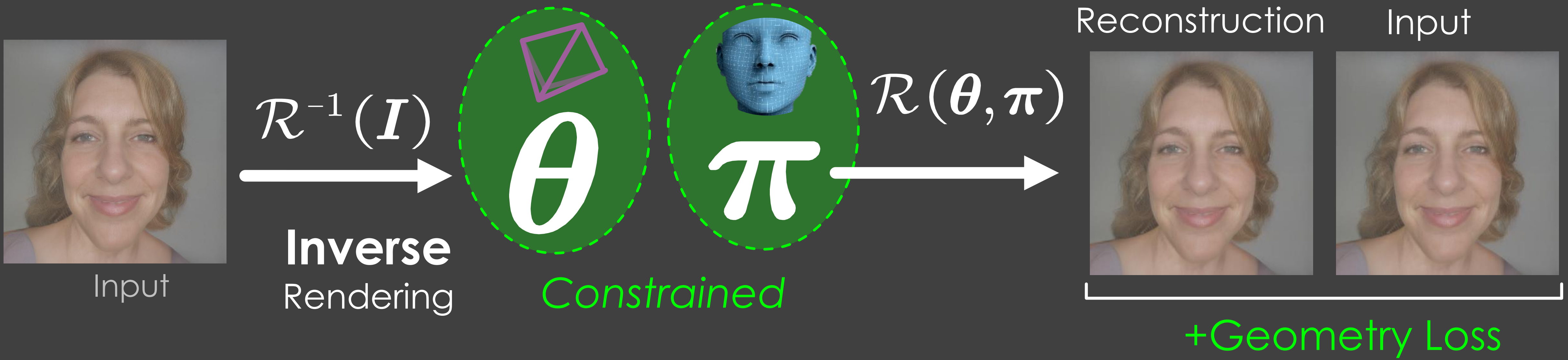
Beyond Matting

- Other ***image-like*** intermediate parameters without accurate label / real date
  - Single Image Normal Map (Single Image)
  - Albedo (Single Image)
  - Depth Estimation (Single Image)

marigold, CVPR’24

Factorization-based Methods

Optimization-based: no labels required



Input

$\mathcal{R}^{-1}(\boldsymbol{I})$

Inverse Rendering

$\boldsymbol{\theta}$ $\boldsymbol{\pi}$

Constrained

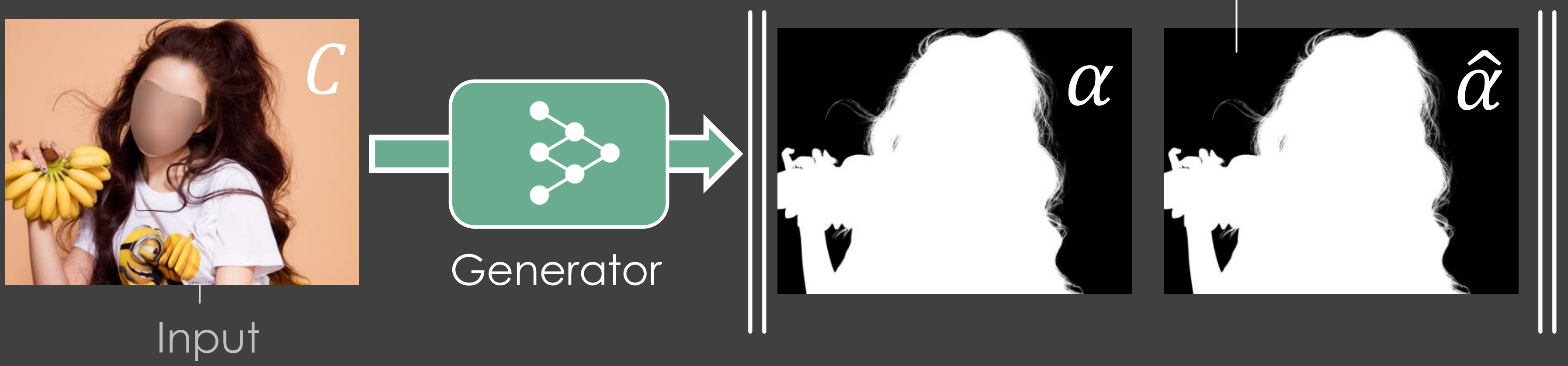
$\mathcal{R}(\boldsymbol{\theta},\boldsymbol{\pi})$

Reconstruction Input

+Geometry Loss

Learning Factorization with Labels: imperfect labels

Human Annotations



C

Generator

$\left\| \alpha \quad \hat{\alpha} \right\|$

Input

Thank you!

Questions or Comments?