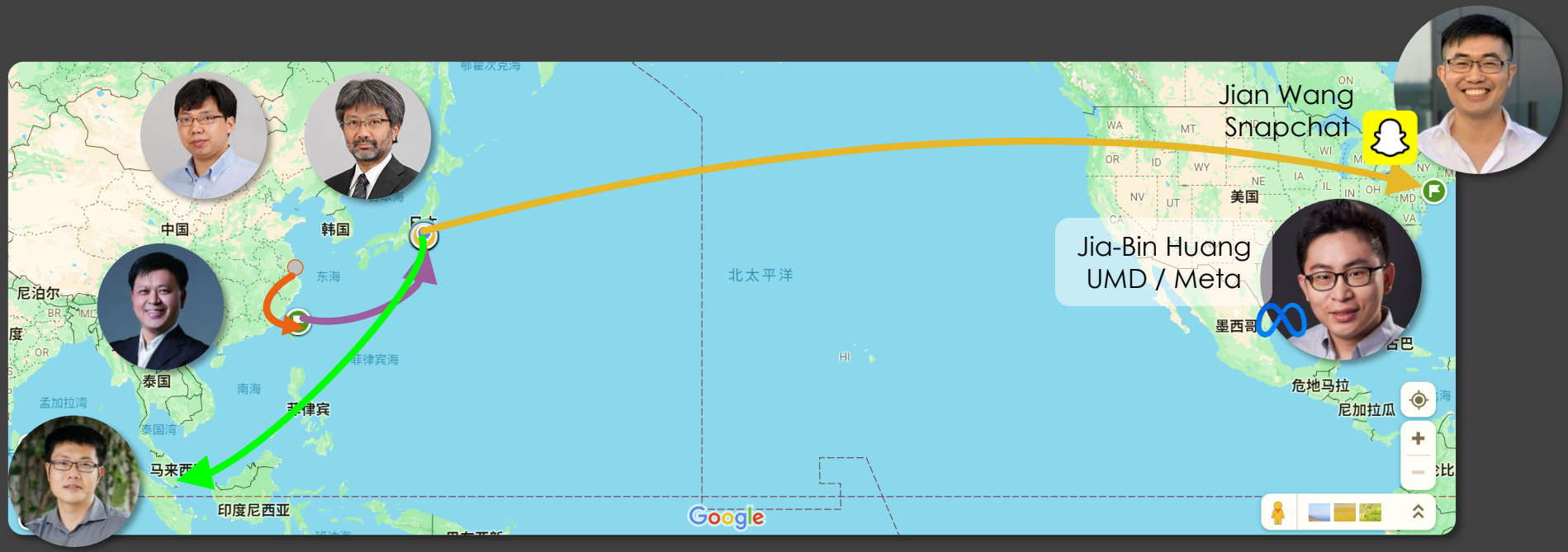
Image Factorization and Manipulation with Generative Regularizations

Zhixiang Wang
PhD candidate
The University of Tokyo

Academic Journey



南京師範大学
Nanjing Normal University

國立臺灣大學
National Taiwan University

東京大学
THE UNIVERSITY OF TOKYO

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

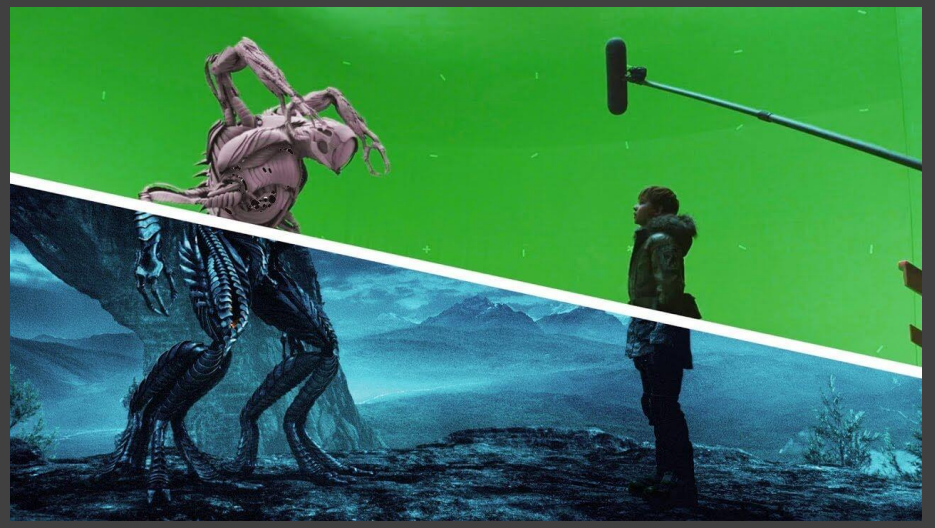
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Snapchat

Sony Research

Goal: GenAI + Advanced Cameras for VFX

Reduce actor, time, and money costs



Research Works

Special Hardwares

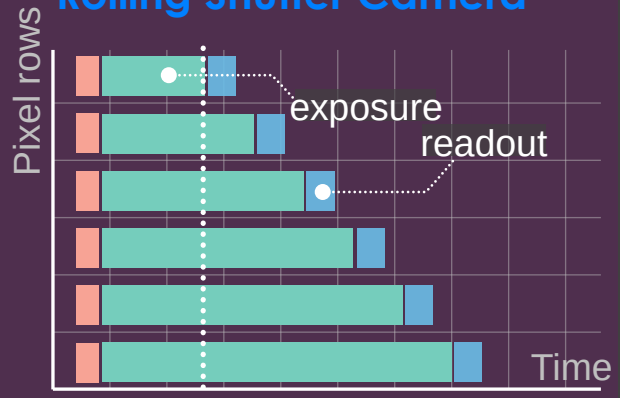
Polarimetric Camera

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

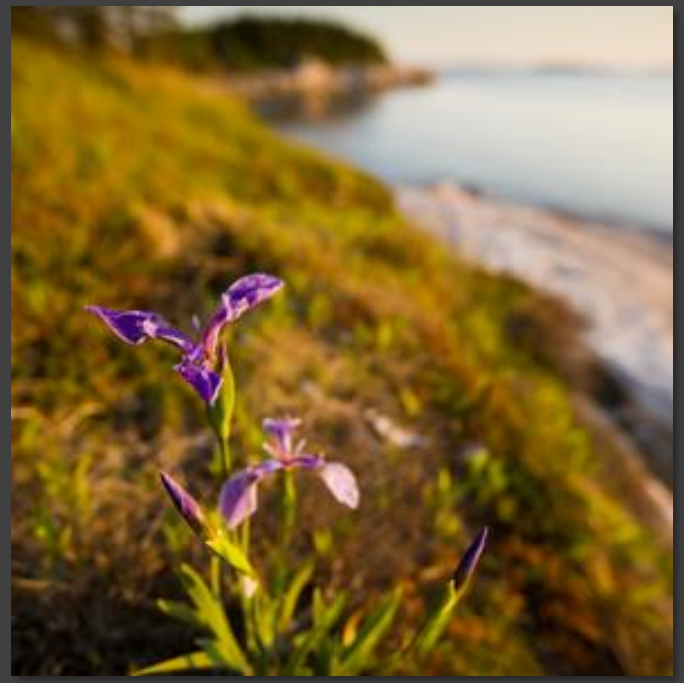
Viewpoint

Background

Difficulty in Controlling Imaging Factors

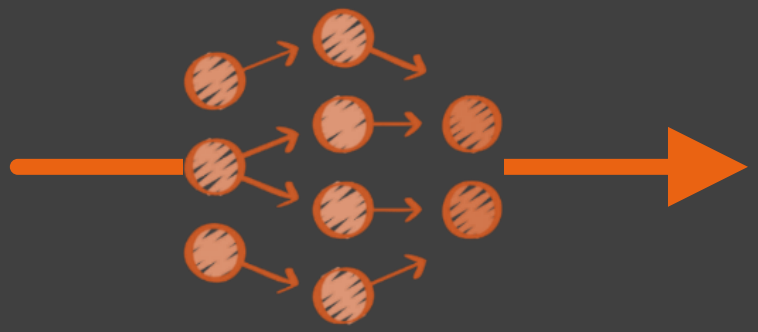
Image credit: DALLE2

Simple yet Popular DL-based Solution



Undesired
Samplings

Image-to-Image Transform

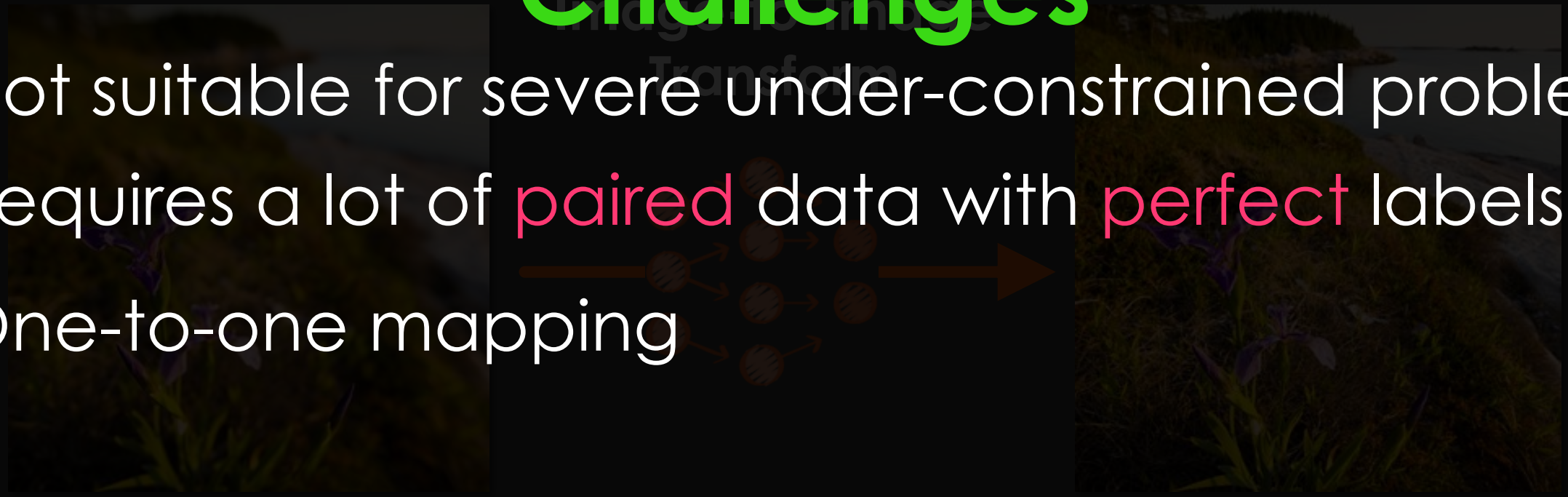


Desired
Samplings

Popular Approaches

Challenges

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



Undesired Samplings

Desired Sampling

Image Factors and Factorization

Undesired Image

Material

Geometry

SCENE

Background

Camera

Viewpoint

Lighting

Motion

Sensor

Lens

Shutter

Aperture

IMAGING FACTORS

Image Manipulation

Undesired Image

Material

Geometry

Scene

Background

Camera

Viewpoint

Lighting

Motion

Sensor

Lens

Shutter

Aperture

Imaging Factors

Render

Re-rendering

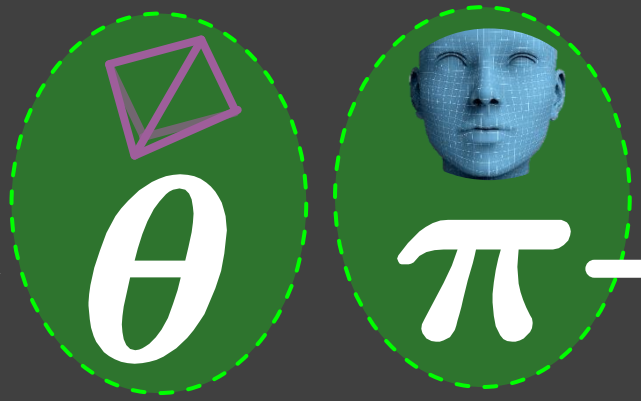
Harness Pre-trained Generative Models

Optimization-based: no labels required

Input

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

Inverse
Rendering



Constrained

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

Reconstruction

Input

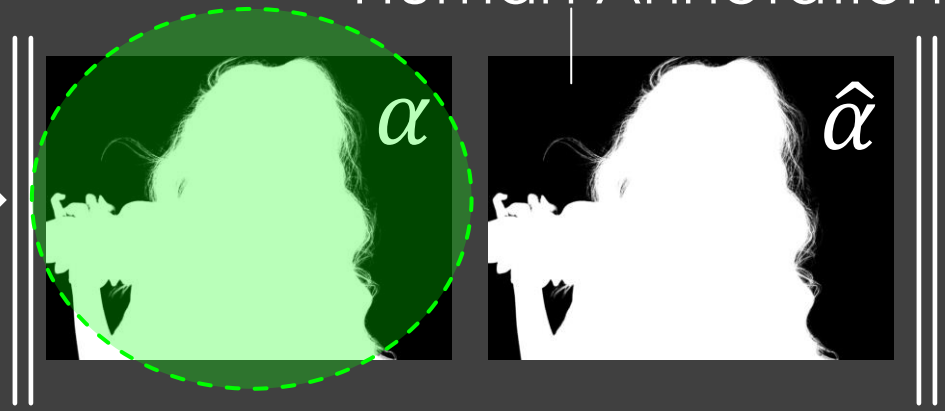
+Geometry Loss

Learning with Labels: imperfect labels

Input

Generator

Human Annotations



Viewpoint + Lens

Background

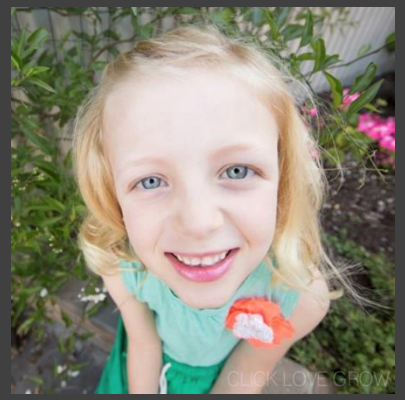
Perspective Distortion Correction

Wang et al, IJCV 2024

Matting by Generation

Wang et al, SIGGRAPH 2024

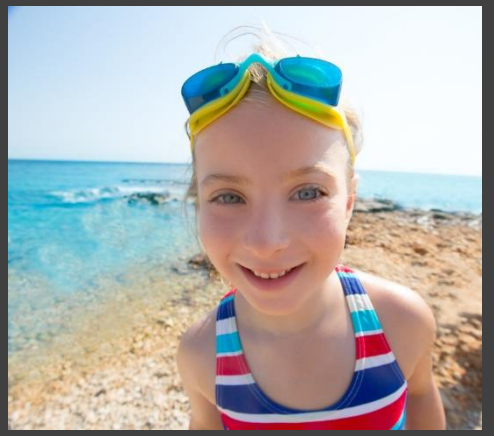




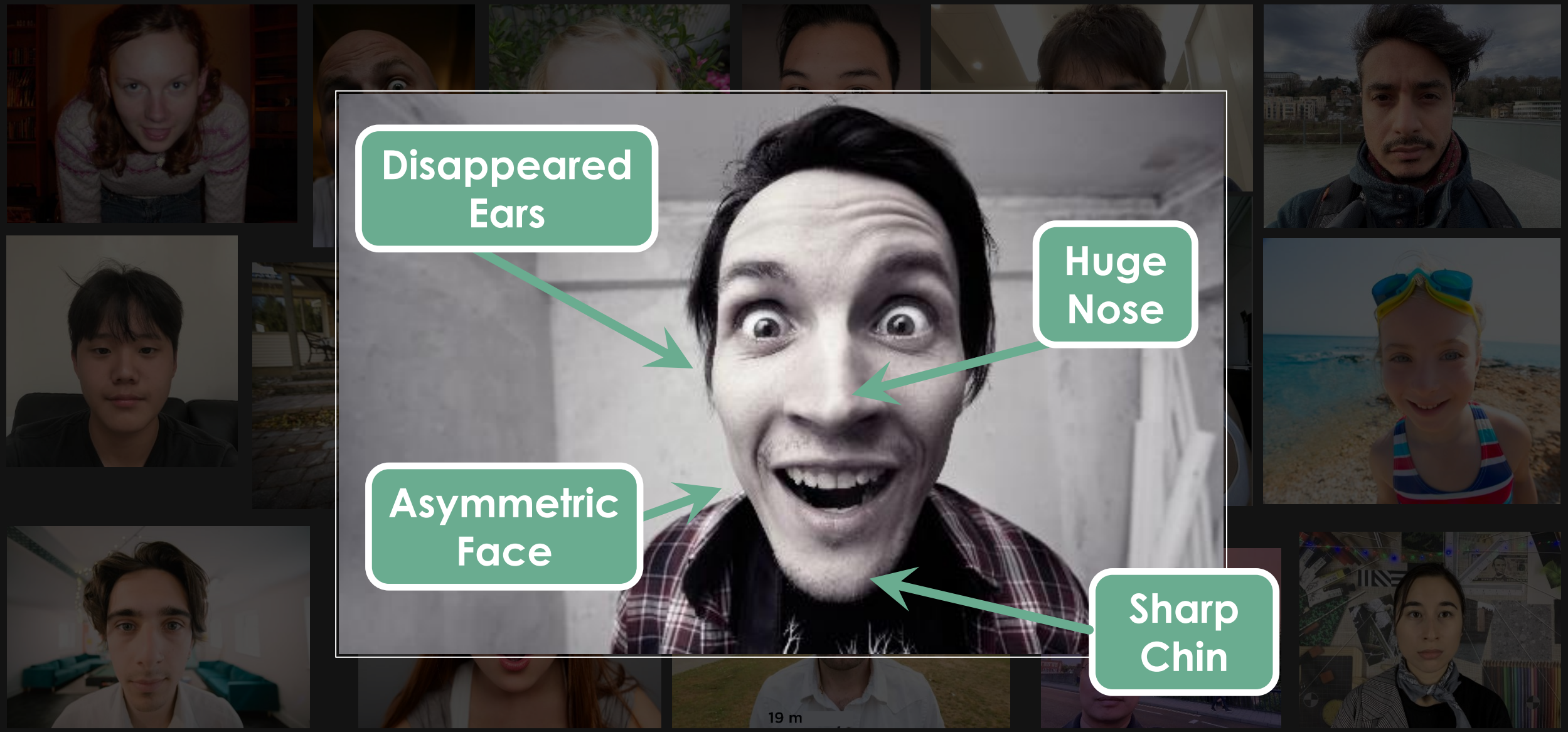












Disappeared Ears
Huge Nose
Asymmetric Face
Sharp Chin

Short Camera-to-Subject Distance

Perspective Projection

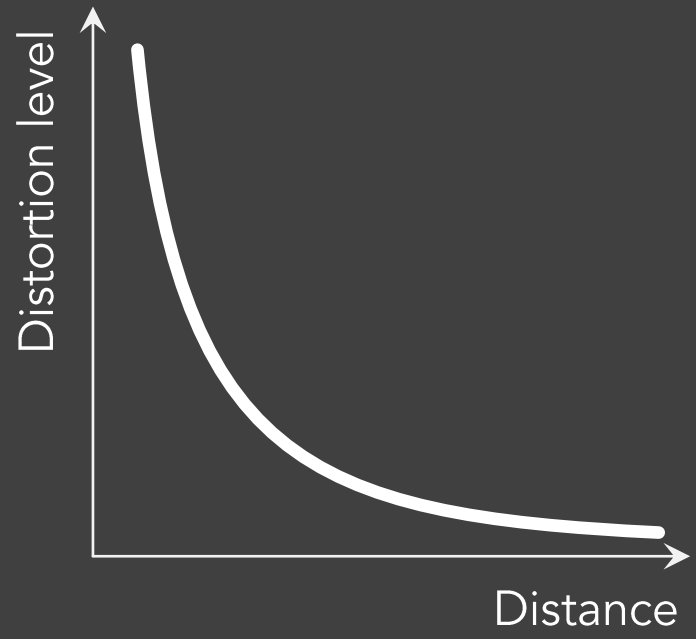
Depth Variation

Δd

D_{close}

Camera-to-Subject Distance

Perspective



Weak-perspective Projection

Perspective

Weak-perspective

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

Depth Variation

Δd

D_{close}

D_{far}

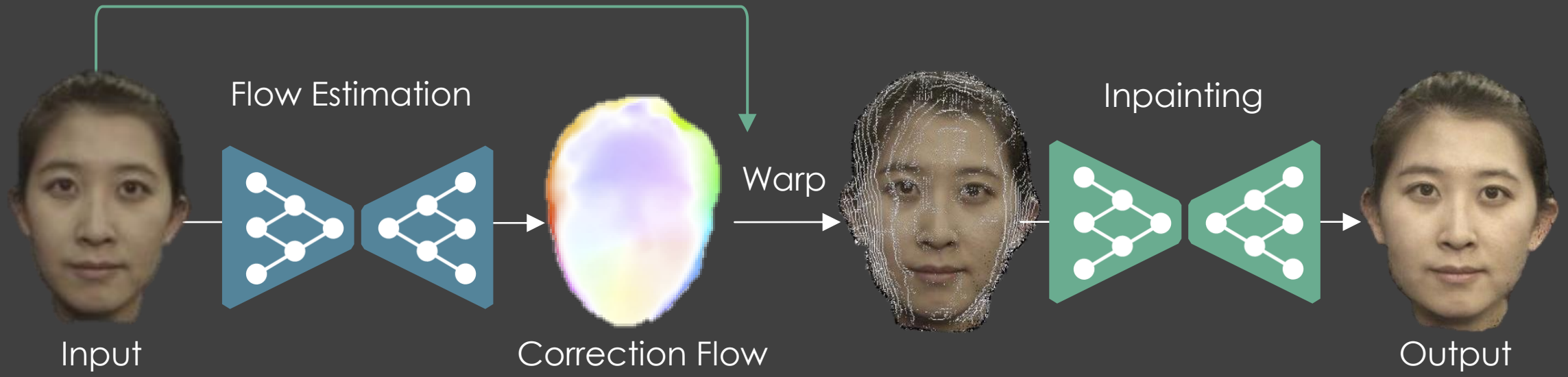
Camera-to-Subject Distance

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

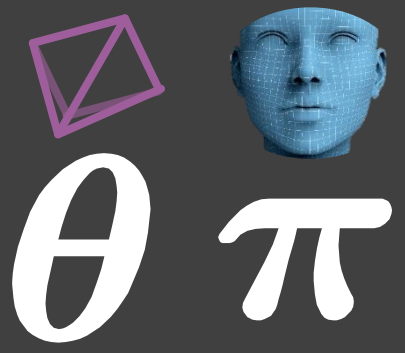
- **Flow warping only repeats existing pixels**
  - CANNOT reveal occluded regions
    - Invisible ear, cheek, neck …
  - CANNOT deal with serious distortion
    - When camera-to-face distance is 20–40cm
  - Not 3D-aware
    - Face shape is flawed
- **Learning-based method (Zhao+) is worse**
  - Require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



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

Forward
Rendering

Reconstruction

Input

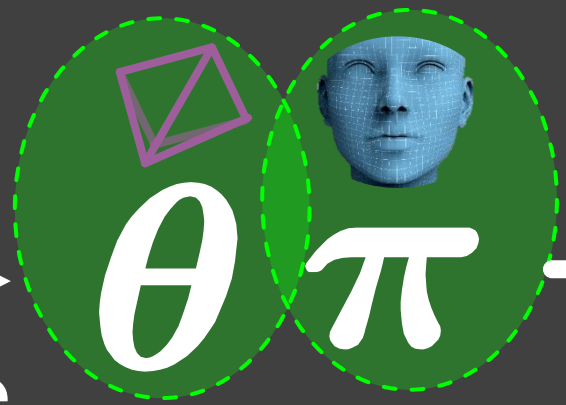
Optimization-based Factorization

Challenge: ill-posed/unconstrained

Input: $\boldsymbol{I}$
Single Image

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

Inverse
Rendering



Constraints

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

Forward
Rendering

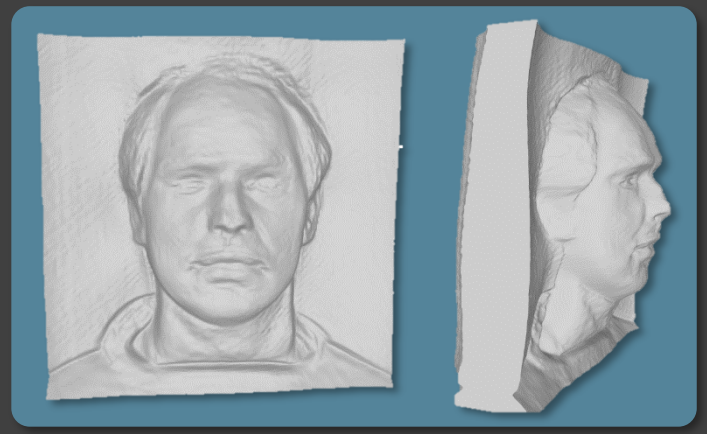
Reconstruction

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

I

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

Inverse Rendering

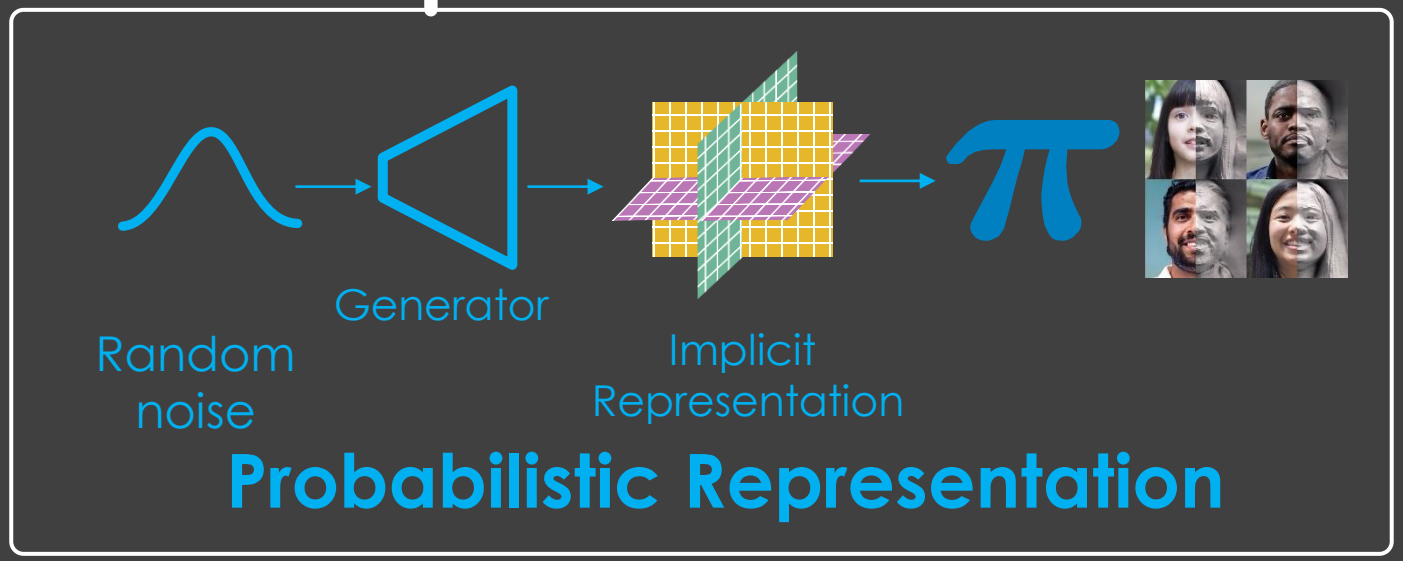
θ π

Unconstrained

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

Reconstruction

3D GAN



Camera Regularization (CR)

Input: $\boldsymbol{I}$
Single Image

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

Inverse
Rendering

$\boldsymbol{\theta}$

Unconstrained

3 Strategies

$\boldsymbol{\pi}$

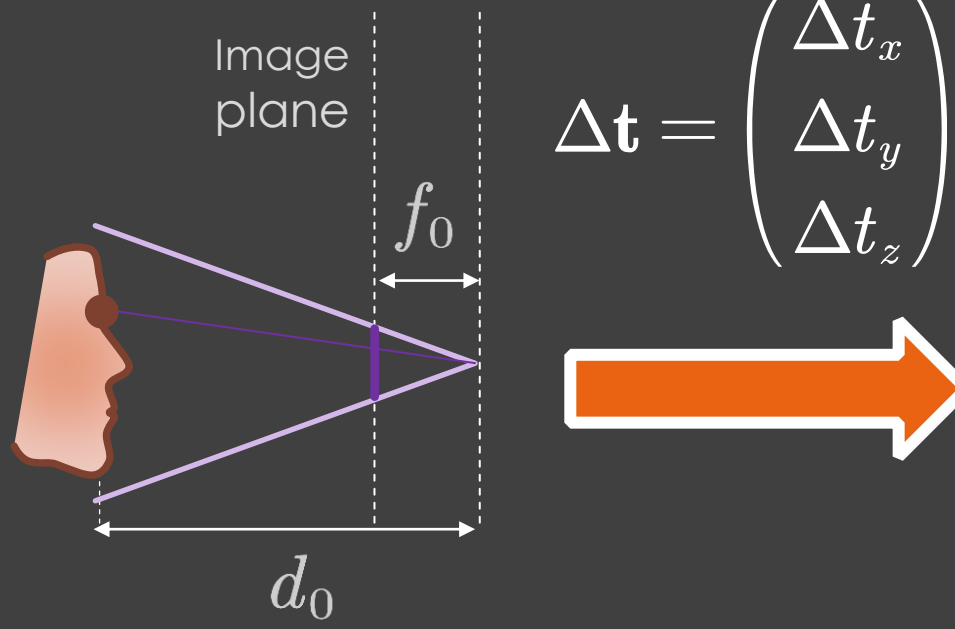
$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{\pi})$

Reconstruction

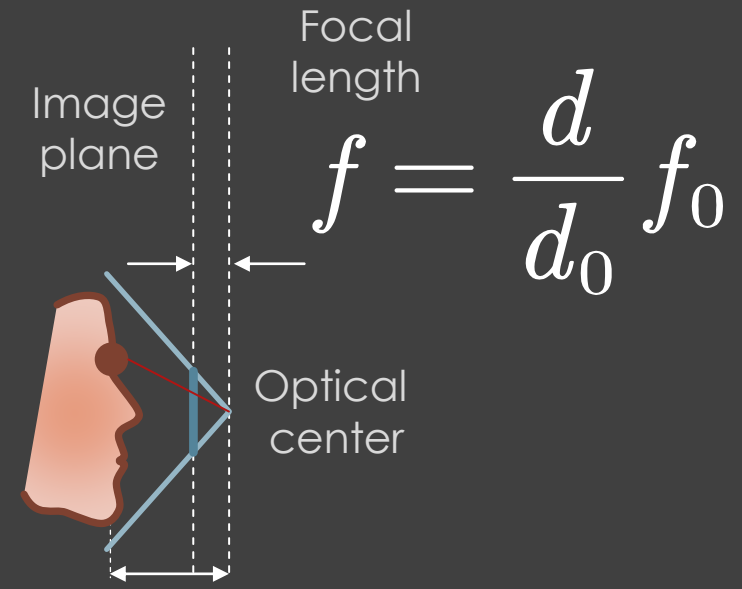
CR 1: Focal Length Re-parameterization

- Focal Length

(simplified approximation)



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

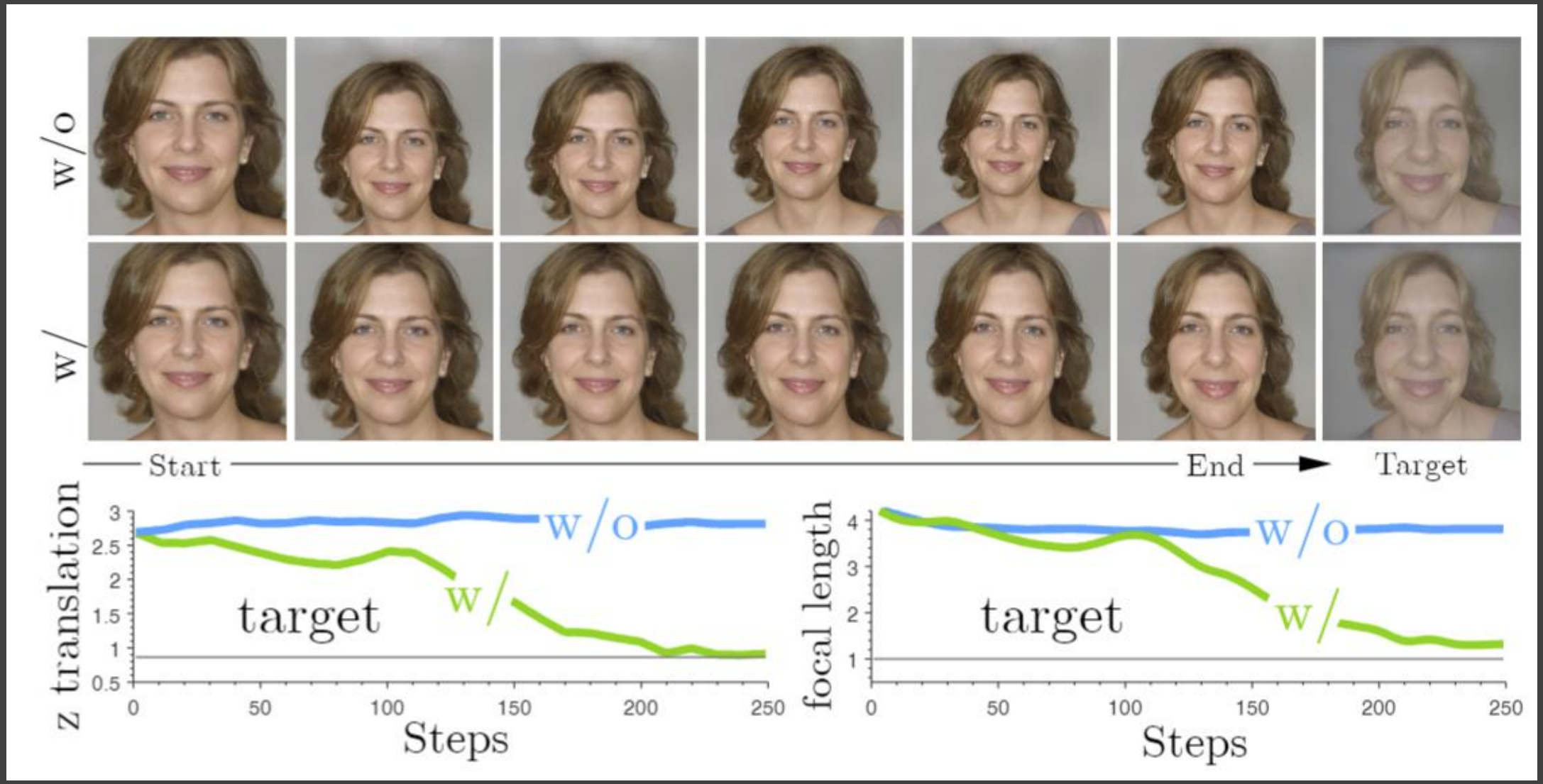


$$f = \frac{d}{d_0} f_0$$

$$d = d_0 + \Delta t_z$$

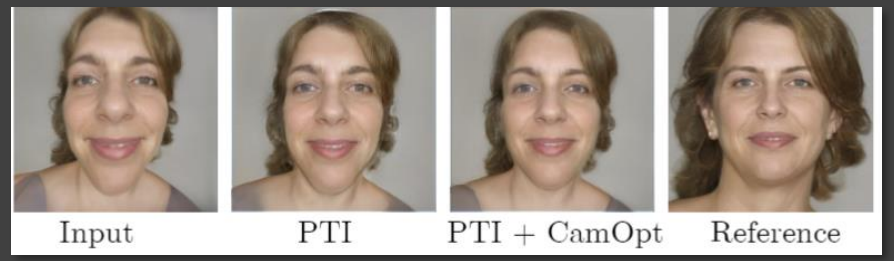
Motivation: Reduce unknown parameters and decouple

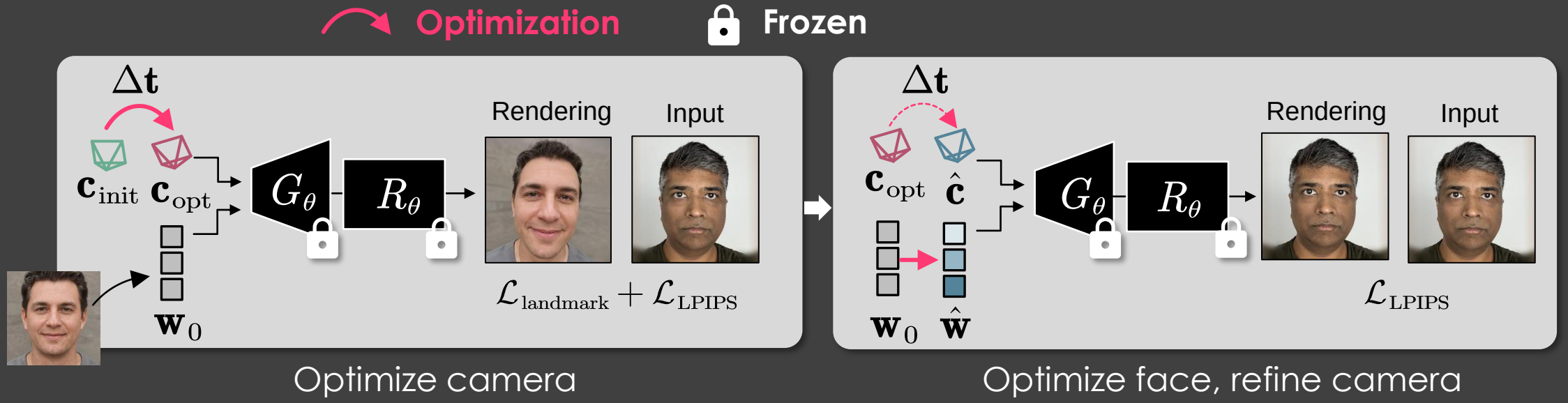
CR 1: Focal Length Re-parameterization



CR 2: Optimization Scheduling

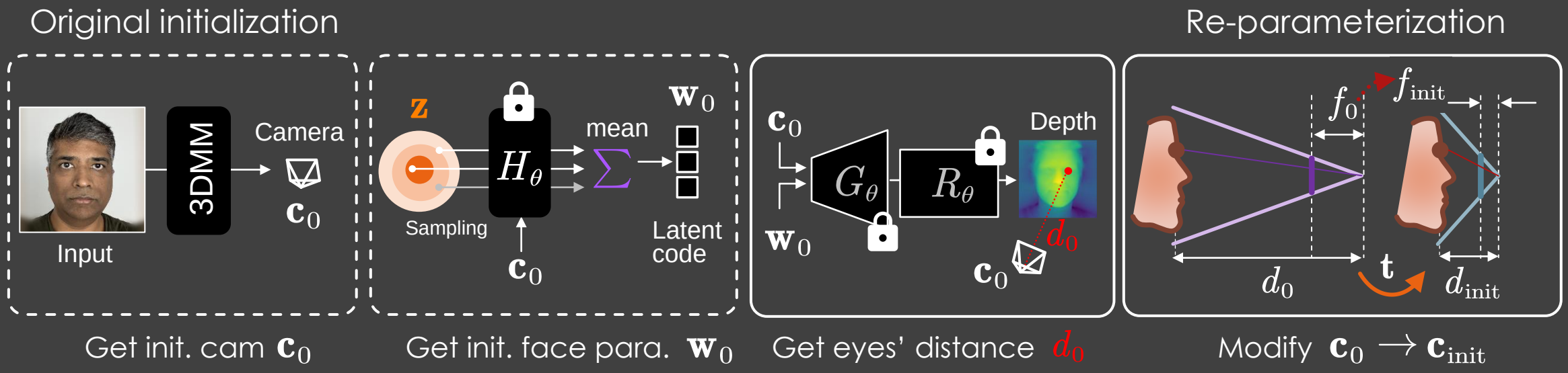
Motivation: Face is **easier** to fall into **sub-optimum** than camera





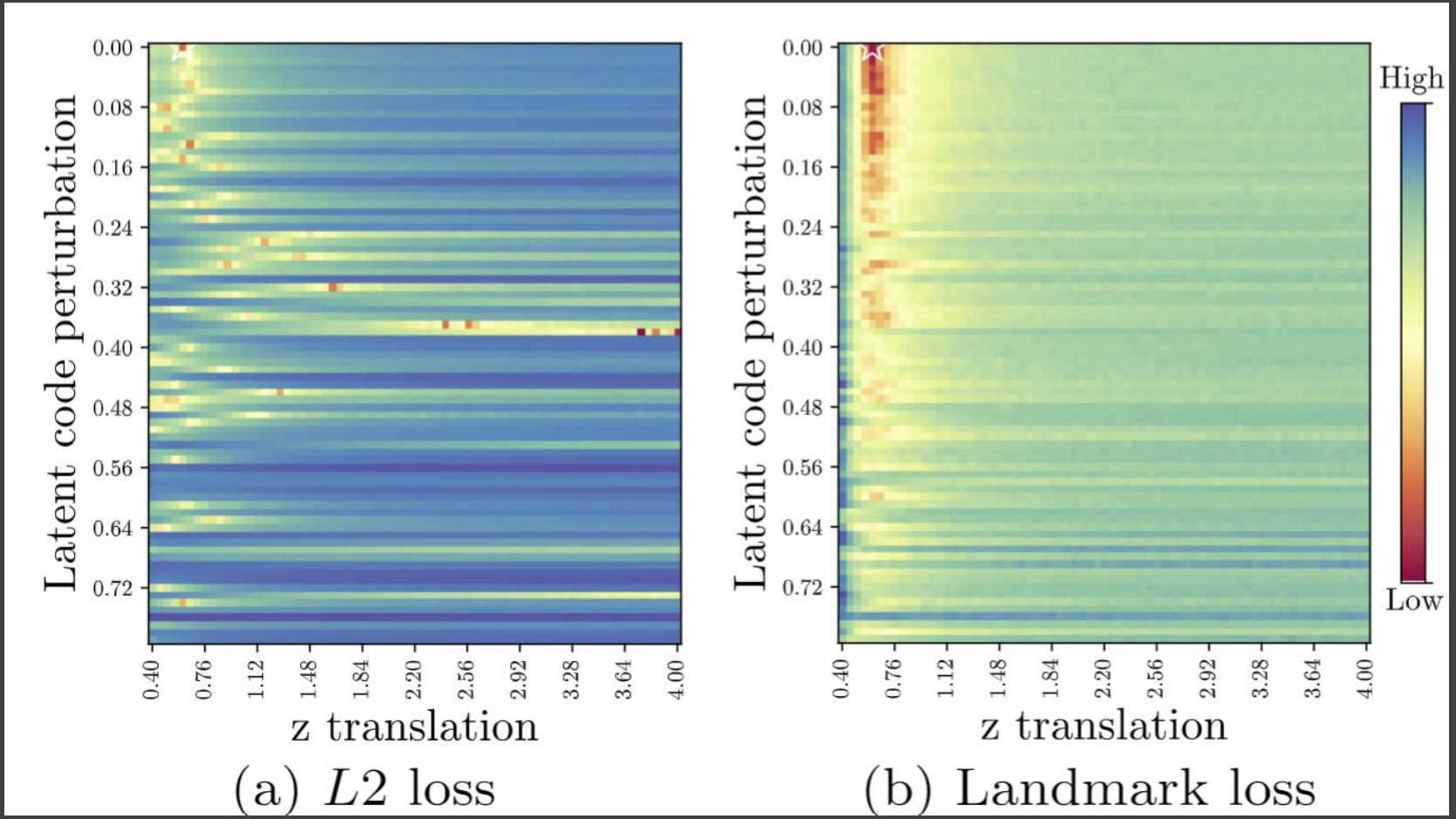
CR 3: Better Initialization

Start from a close-up camera position



Ambiguity Caused by Loss

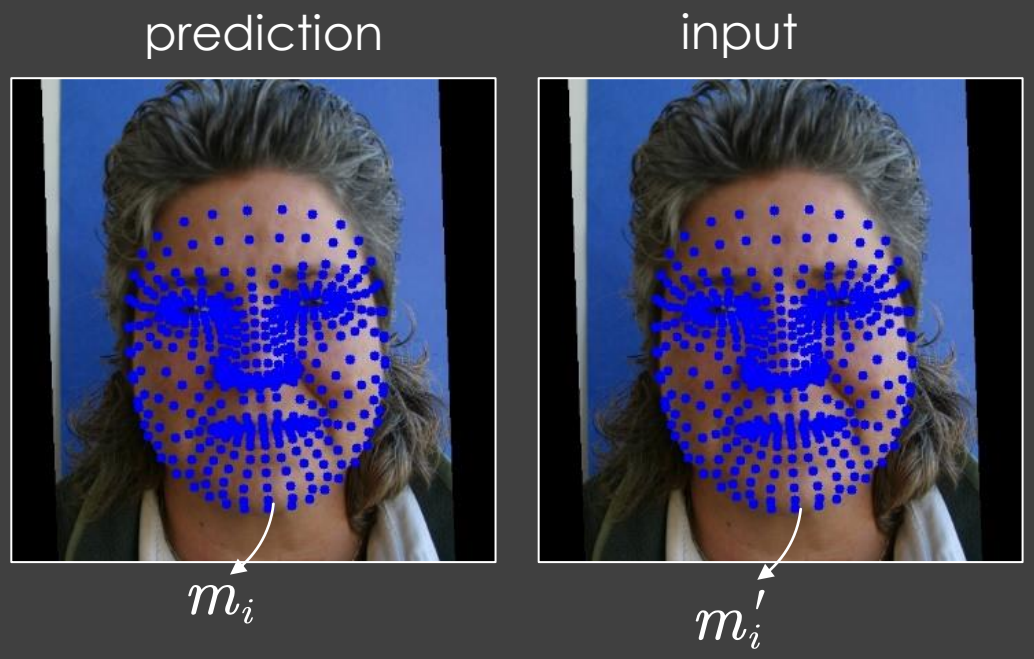
Pixel loss is **very sensitive** to pixel change



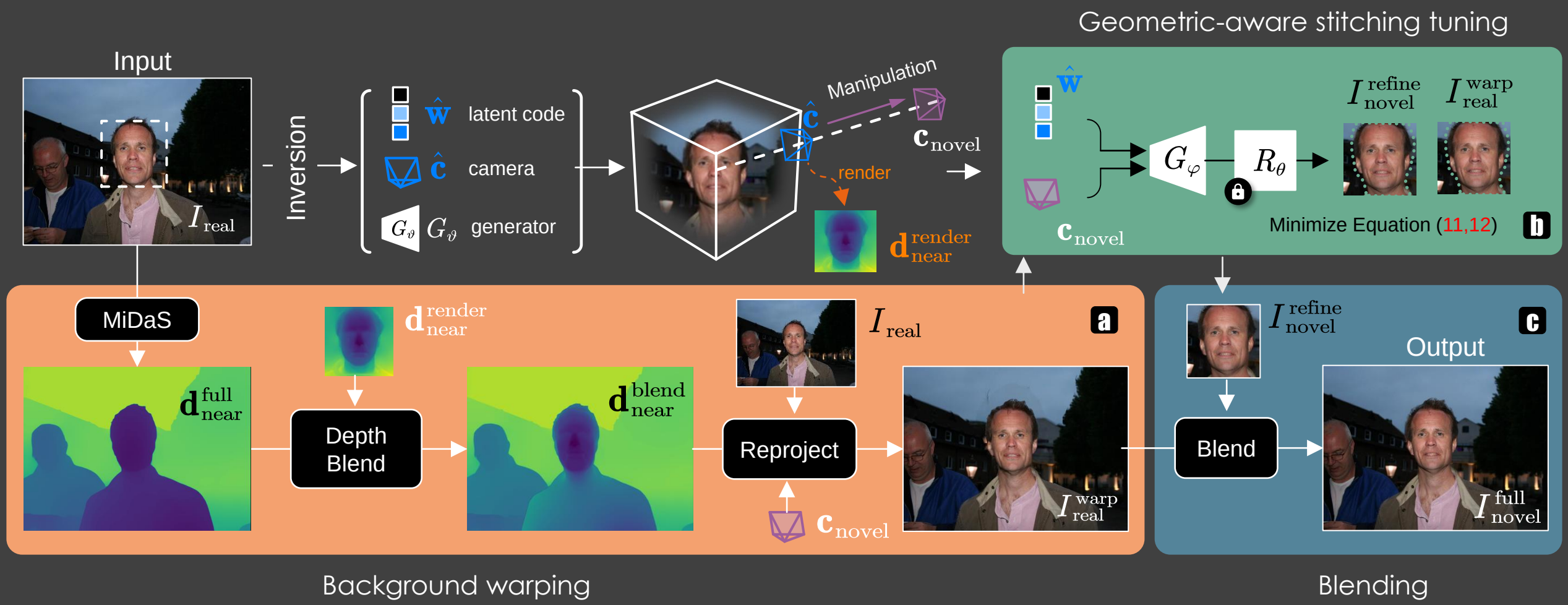
Geometric Regularization

Uncertainty-based Loss

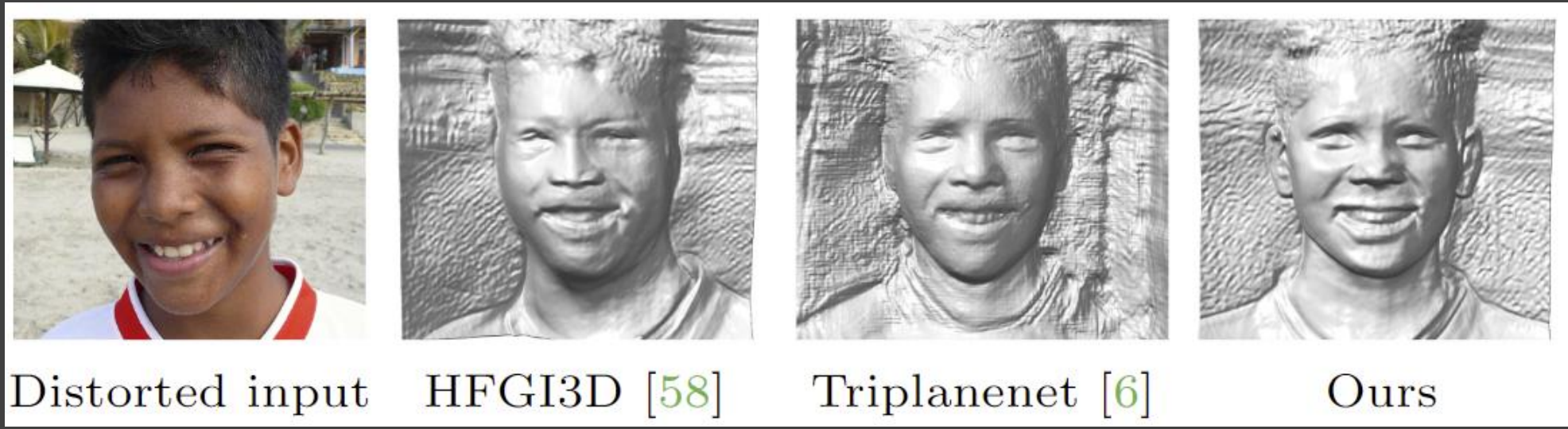
$$\sum_{i=1}^{\|\mathcal{M}\|}\left(\underbrace{\log\left(\sigma_i^2\right)}_{\textbf{Uncertainty term}} + \frac{\|m_i - m_i'\|_2^2}{2\sigma_i^2}\right)$$



Extensions for Full-frame Image



Results – Mesh



Other GAN inversion methods

Results

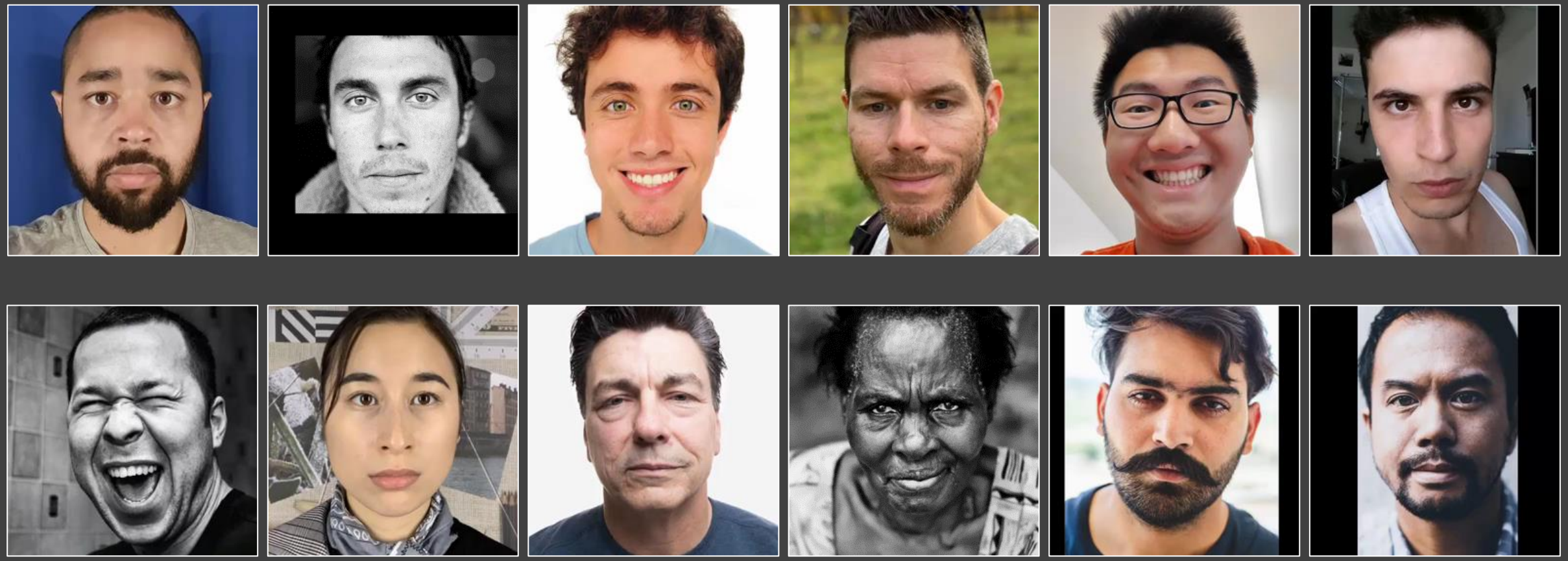
Input




Output




Results – Continuous Manipulation



Results – Comparison

Stretch-like

Fried et al, SIGGRAPH'16

3D **geometric consistent**

Ours

Results – Comparison

Input

Fried et al, SIGGRAPH'16

Ours

Results – Comparison

Input

Fried et al, SIGGRAPH'16

Ours

Results – Comparison

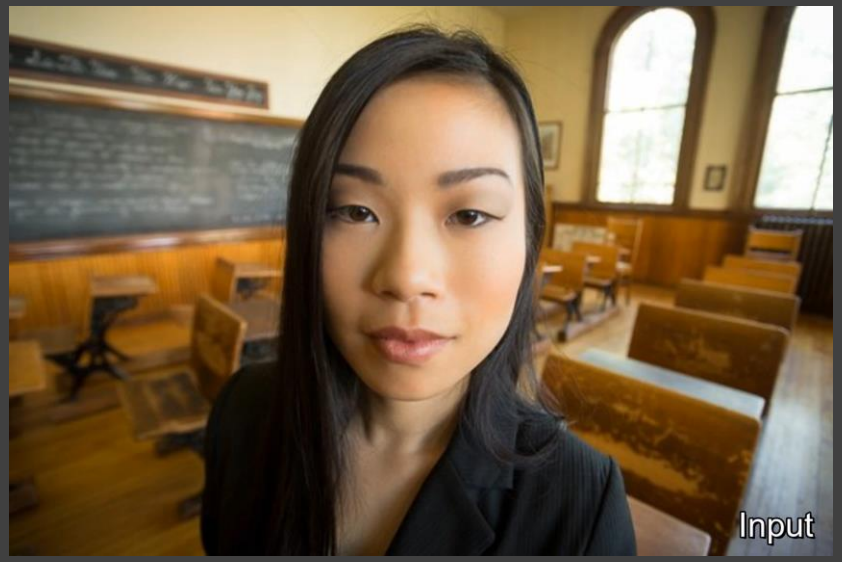
Input

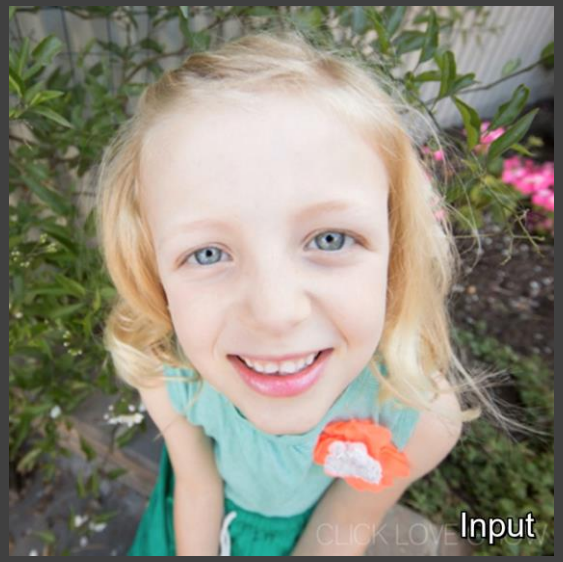
Fried et al, SIGGRAPH'16

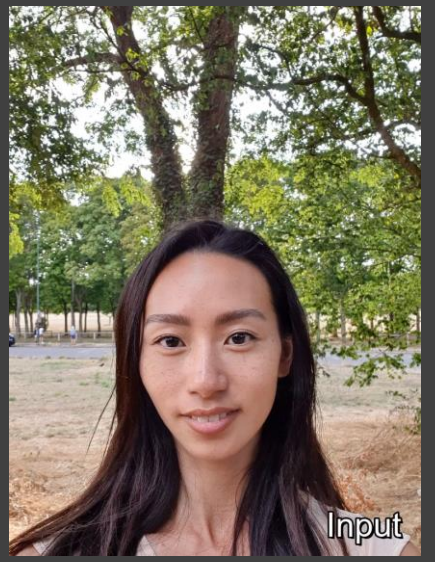
Ours

Dolly Zoom

Dolly Zoom







Viewpoint + Lens

Perspective Distortion Correction

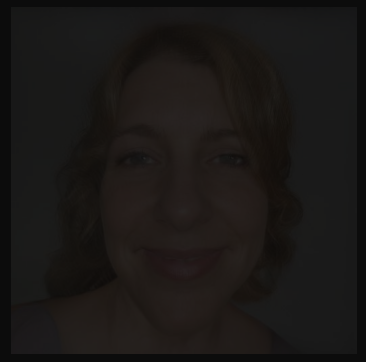
Background

Matting by Generation

Wang et al, SIGGRAPH 2024

Harness Pre-trained Generative Models

Optimization-based: no labels required



Input

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

Inverse
Rendering

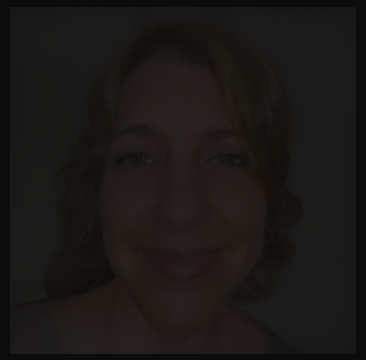
θ π

Constrained

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

Reconstruction

Input



+Geometry Loss

Learning with Labels: imperfect labels

C

Input

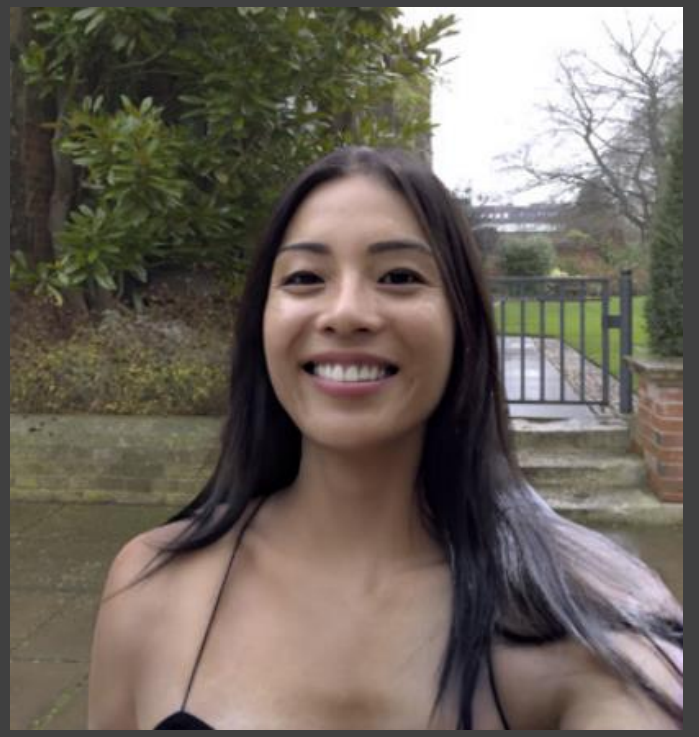
Generator

α

Human Annotations

$\hat{\alpha}$

Manipulate Background



Background Gallery

Image credit: total relighting

Factorization Problem

Input: $\boldsymbol{I}$
Single Image

Matting

Inverse
Rendering

Composition

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

Forward
Rendering

Re-rendering
$\alpha F + (1 - \alpha)B^*$

Learning with Labels

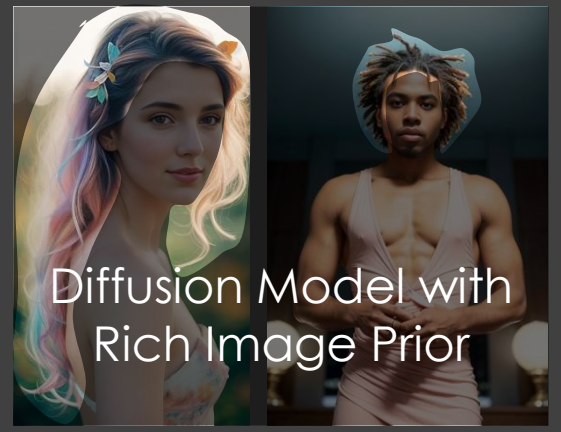
Poor label quality

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

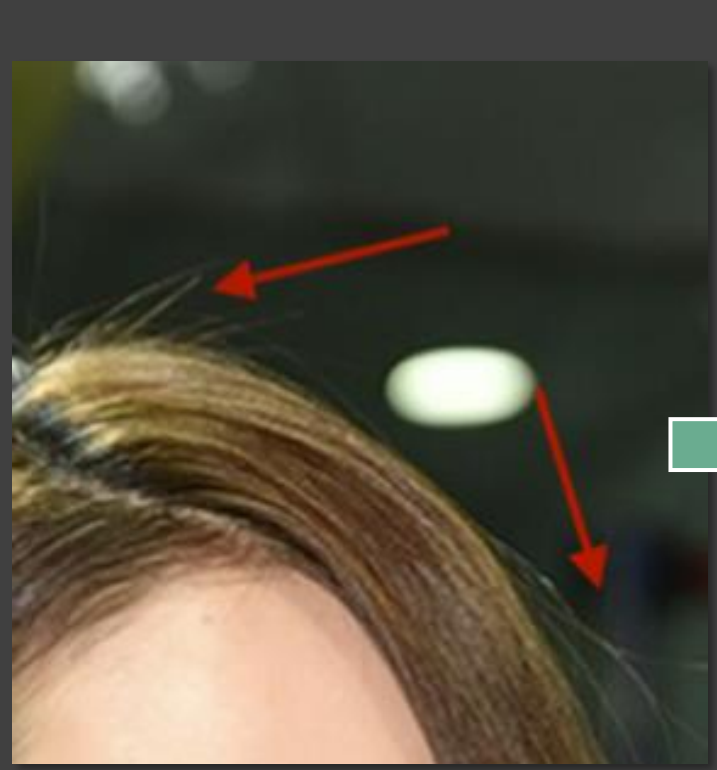
Limitations of Existing Methods

Ke et al, MODNet, AAAI'22

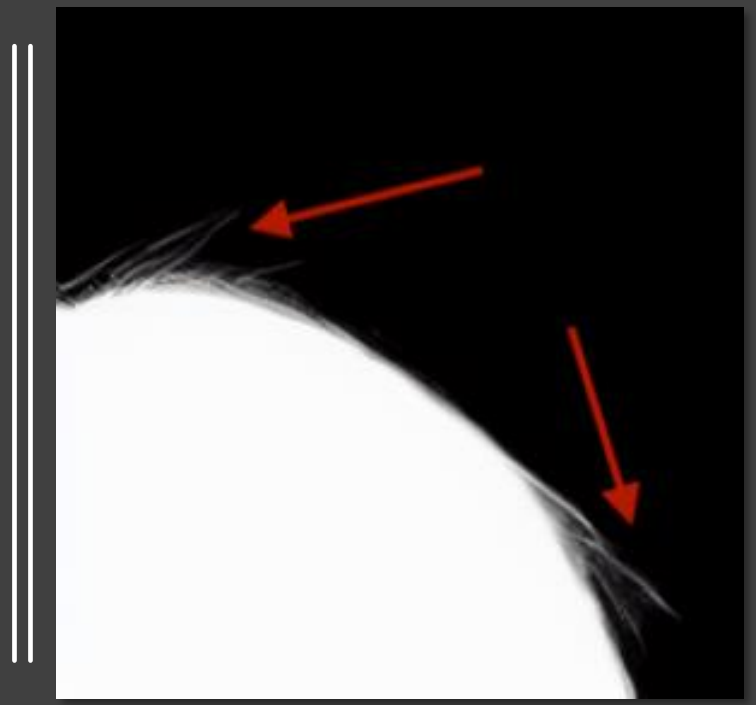
Generative Diffusion Prior



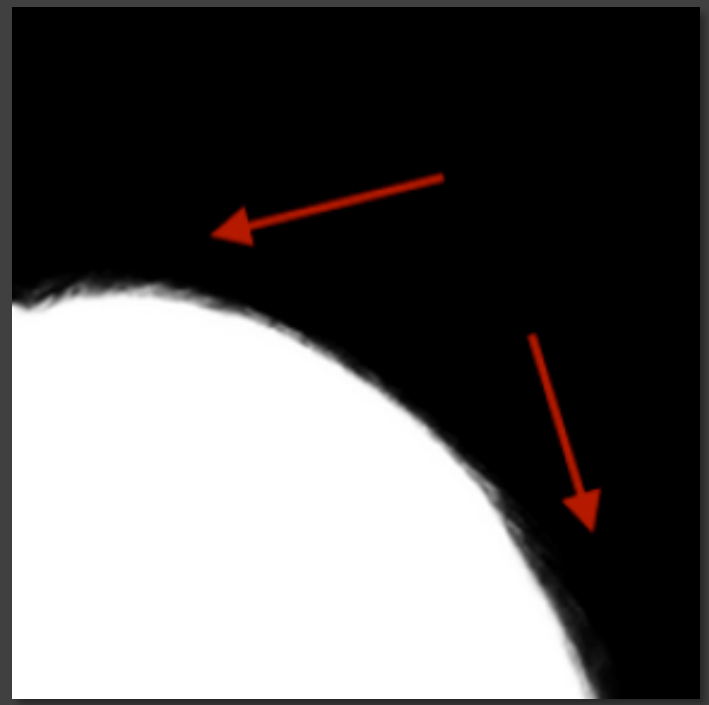
Generative Prior for Regularization



Input training image

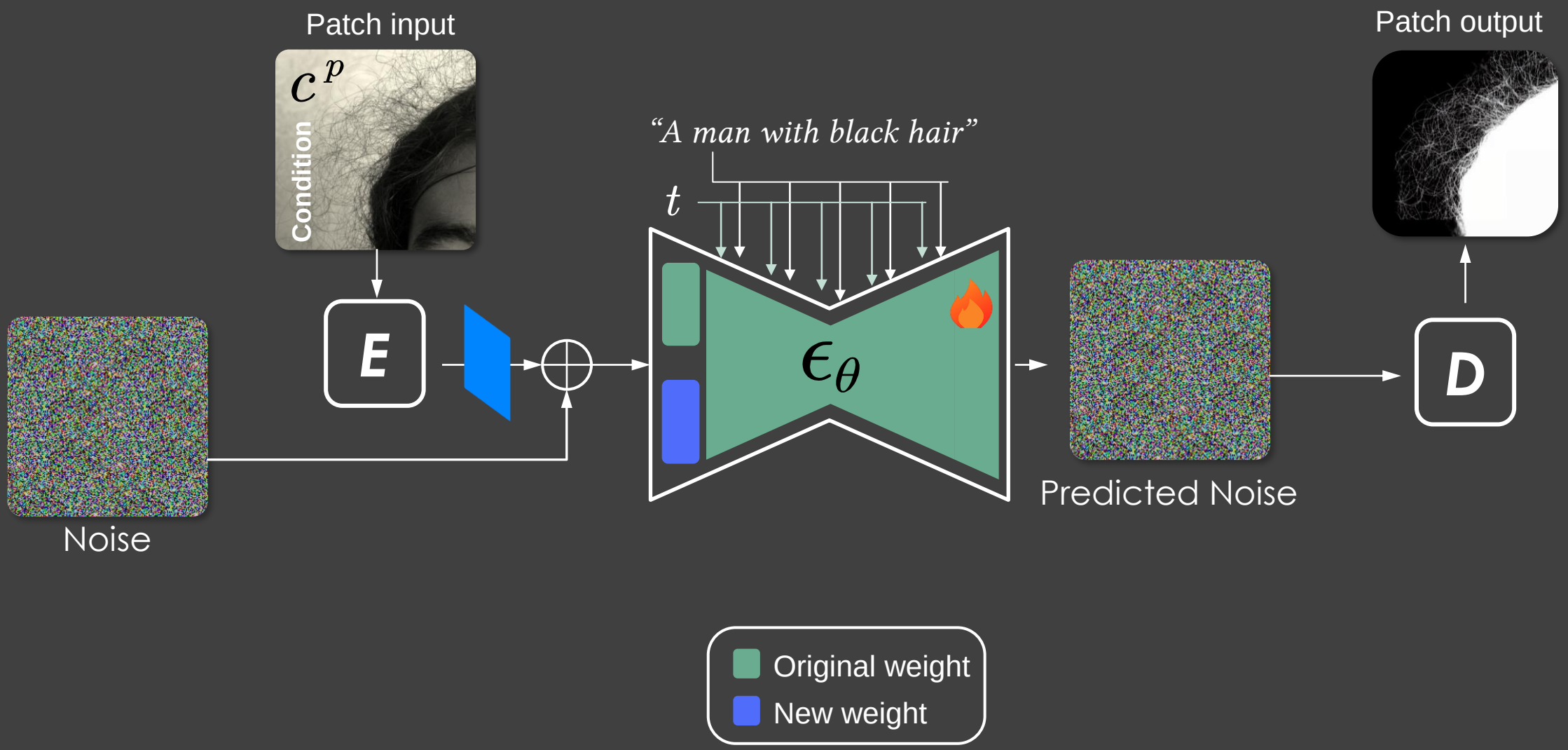


Output

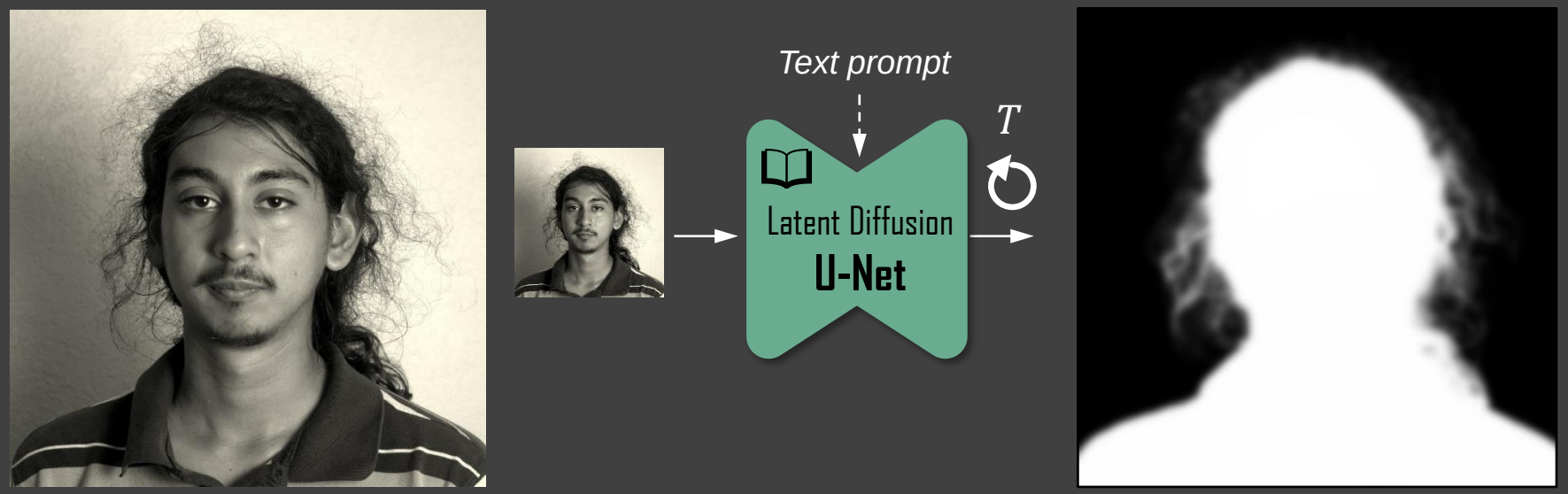


Label

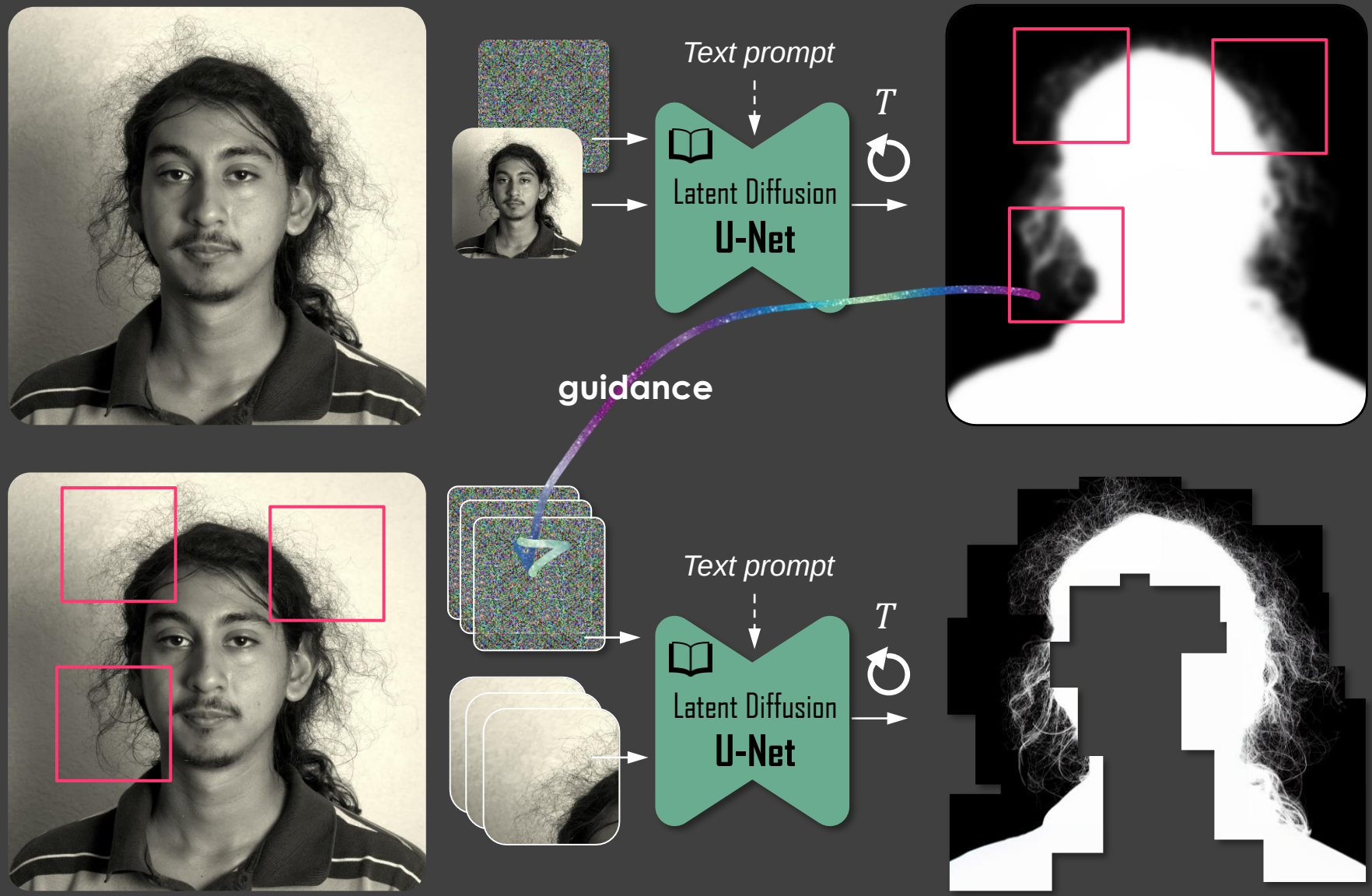
Repurposing Latent Diffusion Model



Challenge of Processing HR Images



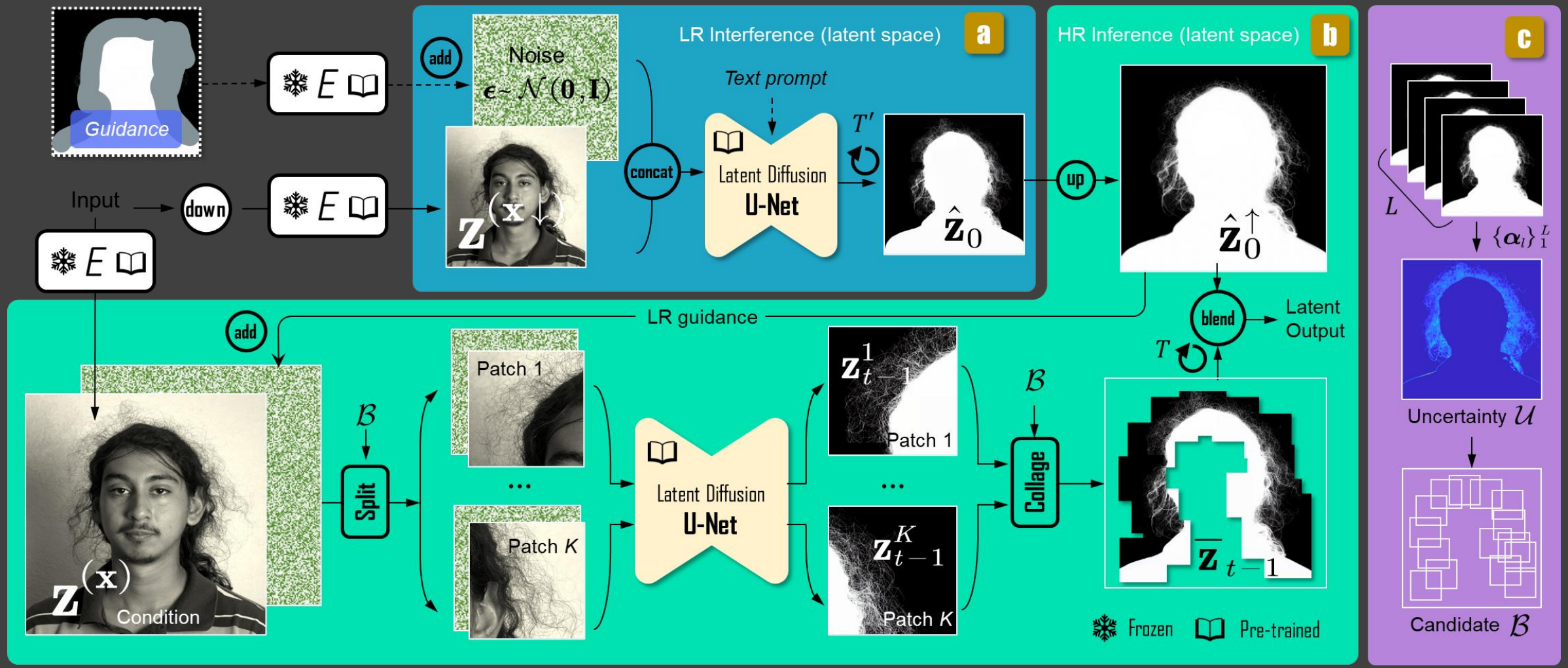
Pipeline for Processing HR Images

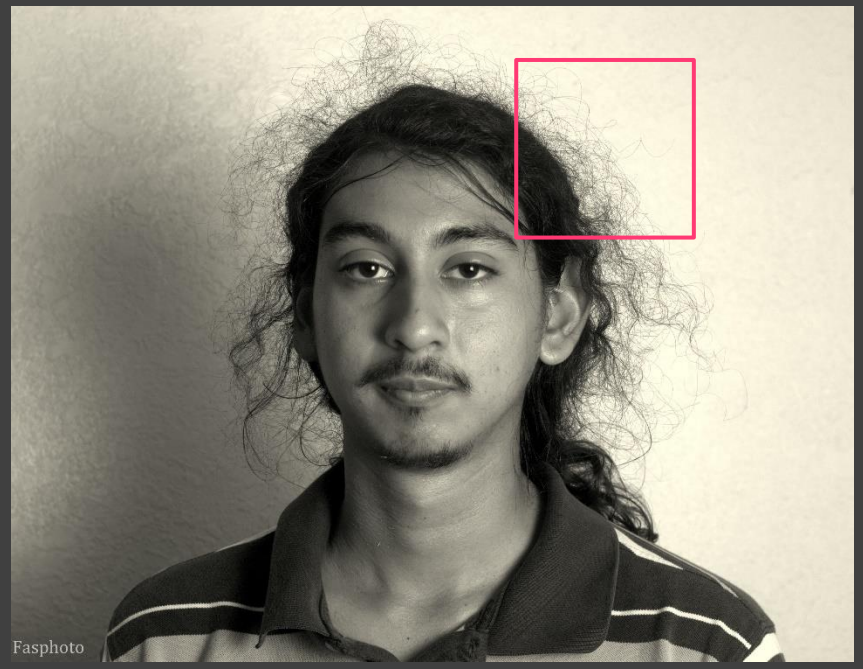


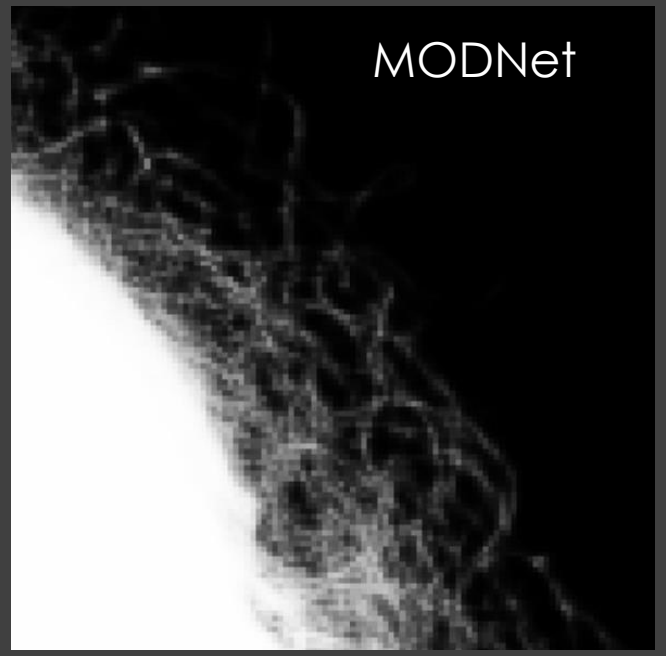
Pipeline for Processing HR Images

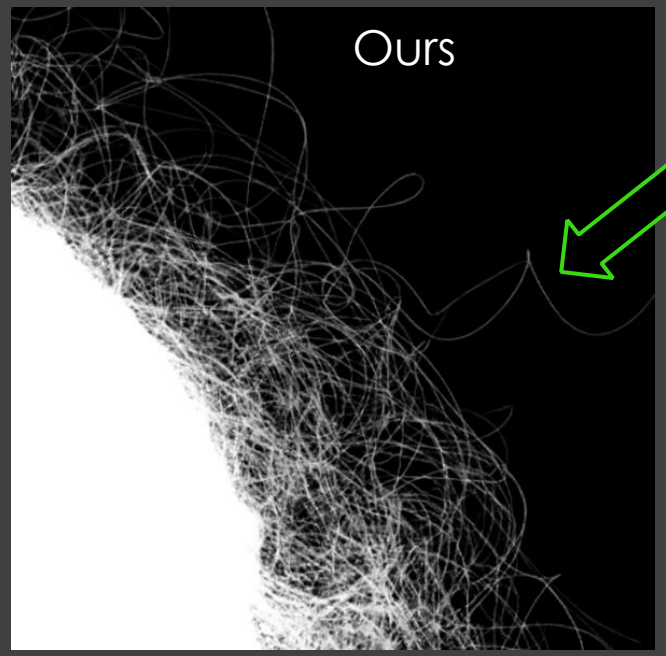
Get potential areas by uncertainty

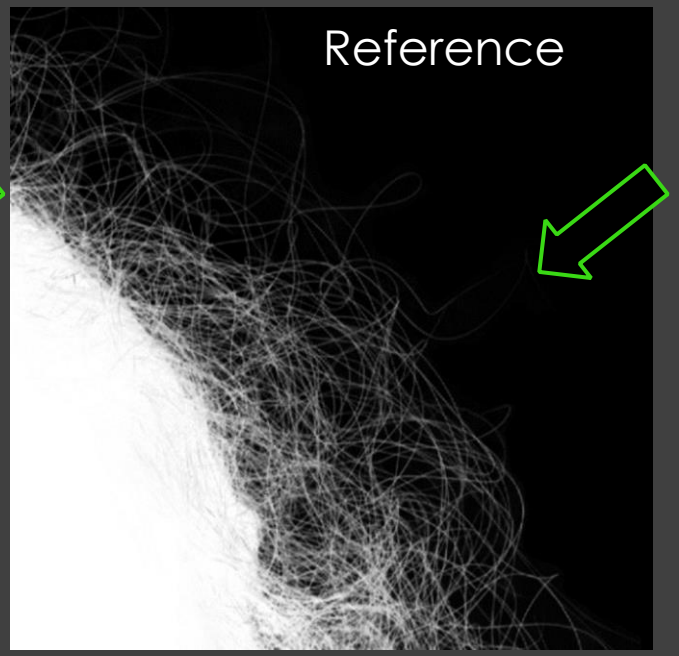
Full Pipeline











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

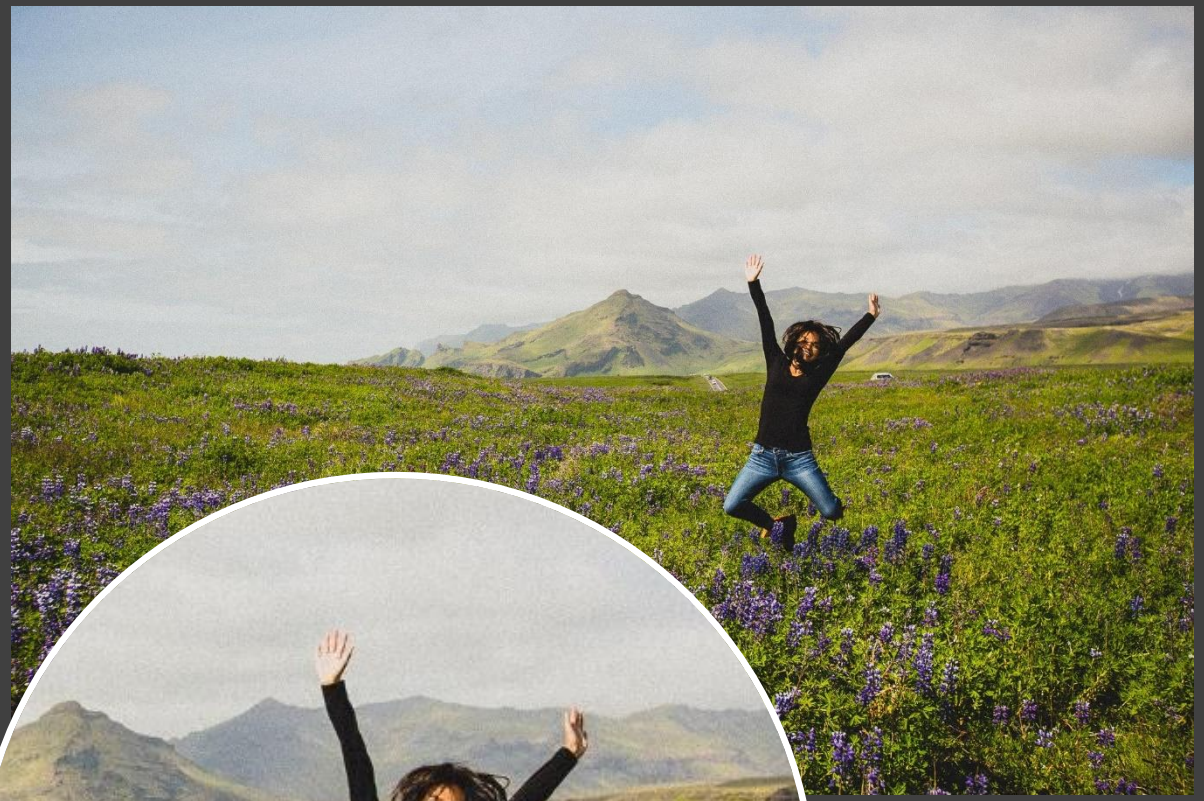
Ours

Input

ViTAE-S

Ours


Input
PUSH BUTTON FOR
ViTAE-S
Ours

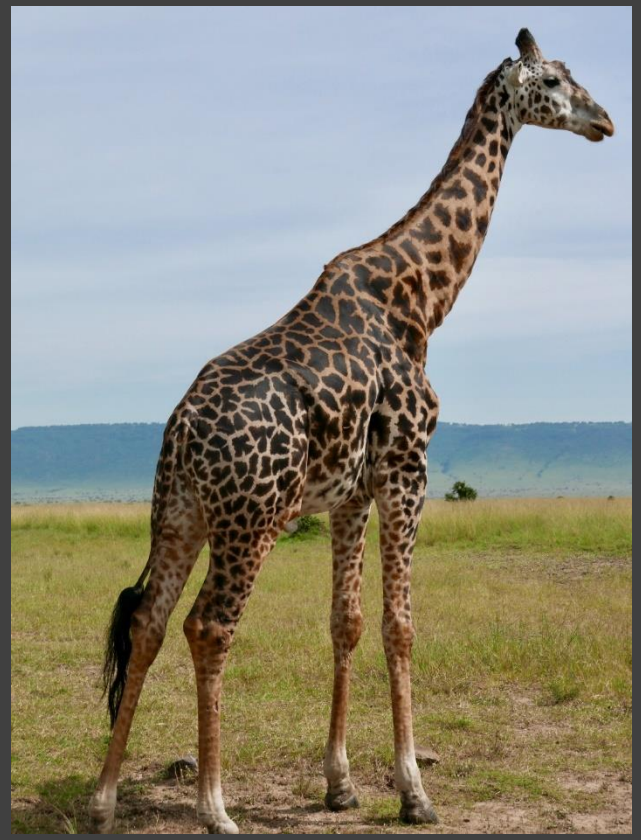


ViTAE-S

Ours

Out-of-Distribution Matting

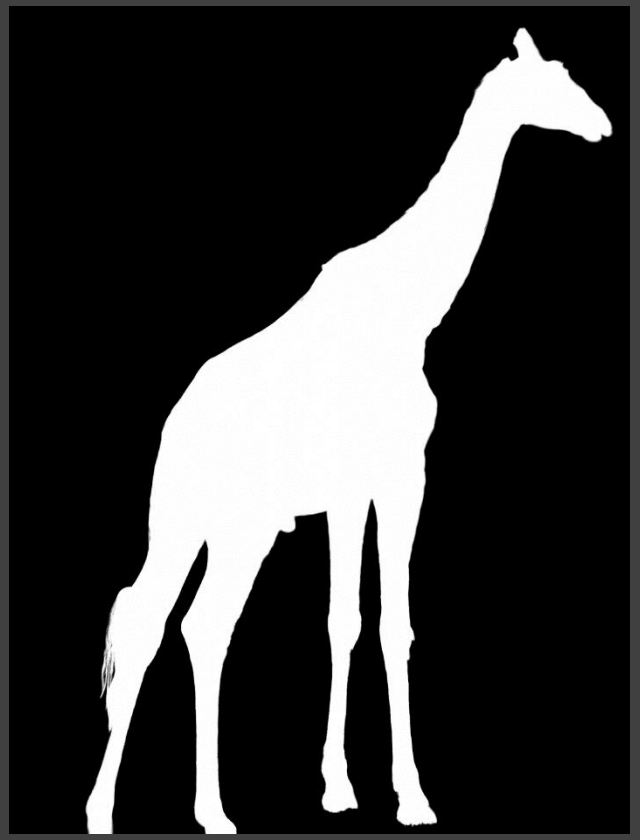
Input



SAM-based

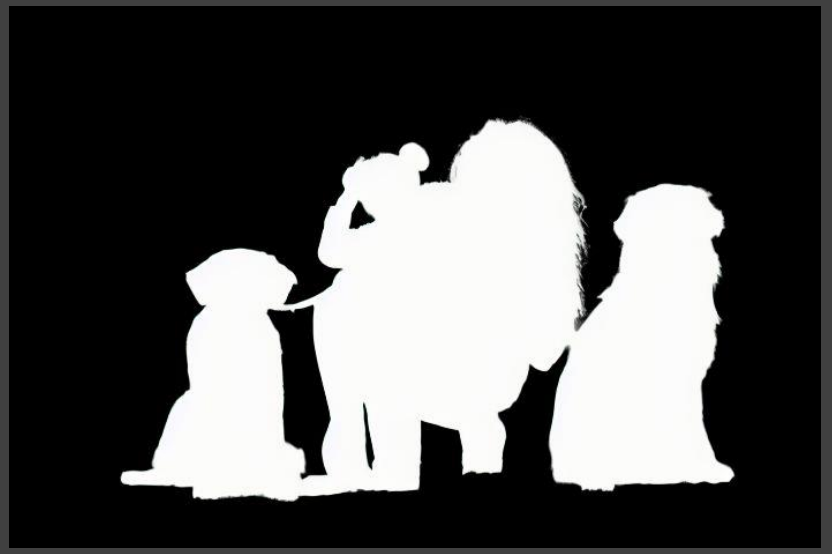
ViTAE-S

Ours



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

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

Inverse
Rendering

θ

π

Constrained

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

Reconstruction

Input

+Geometry Loss

Learning Factorization with Labels: imperfect labels

Input

Generator

Human Annotations

Thank you!
Questions or Comments?