3D Face Estimation from a Monocular RGB Image with Dense Landmarks

Master Thesis

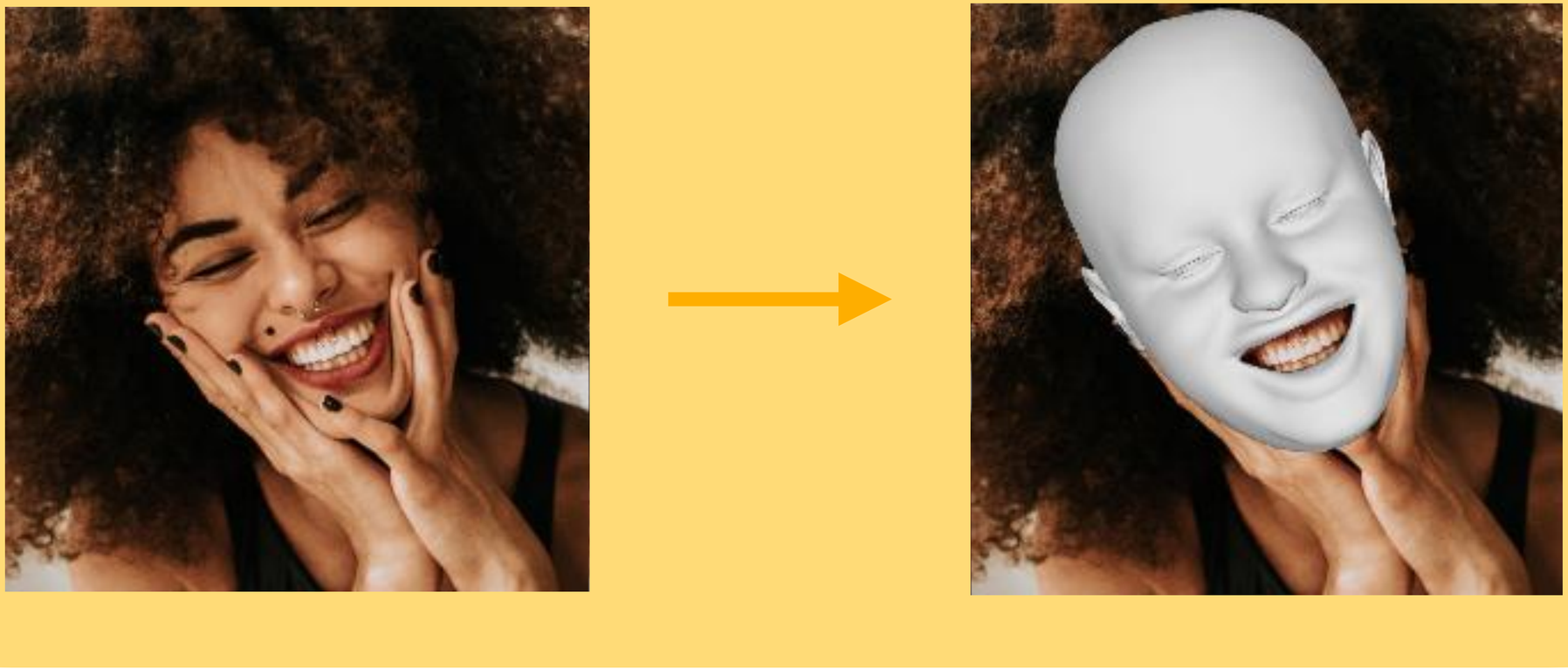
CVL Computer Vision Lab

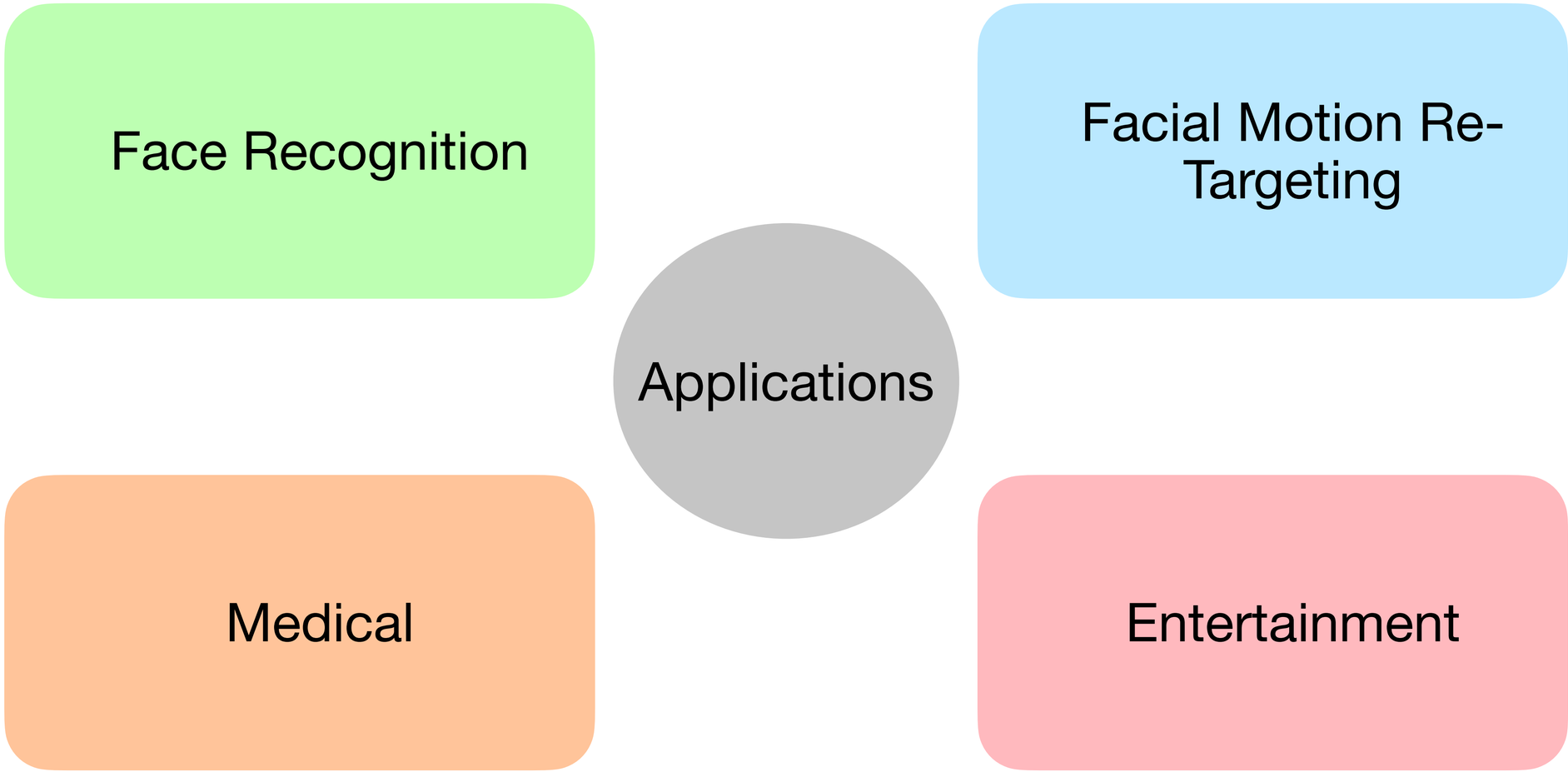
Nicholas Simic

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

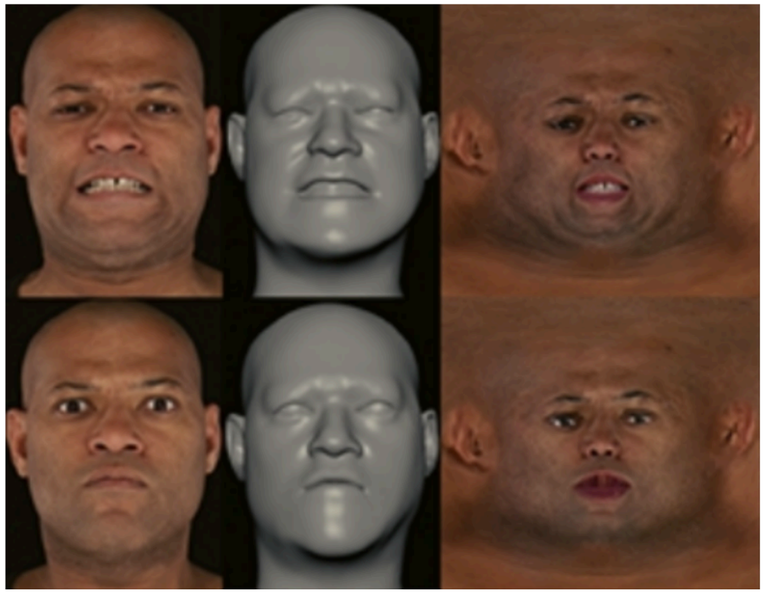
3D Face Mesh Reconstruction

3D Face Mesh Reconstruction




Face Recognition
Facial Motion Re-Targeting
Applications
Medical
Entertainment

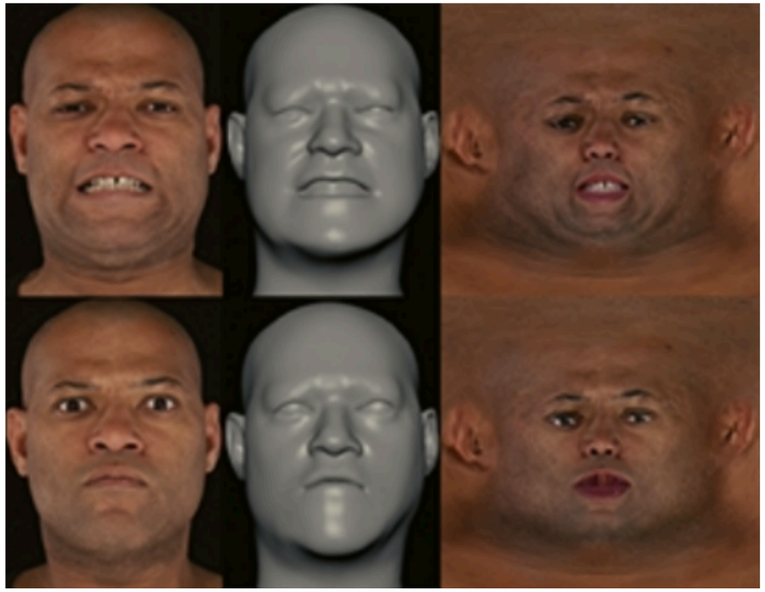
Multi-View and Motion Capture in Movies



Yeongho Seol, Wan-Chun Ma, and John P Lewis. “Creating an actor-specific facial rig from performance capture”. In: *Proceedings of the 2016 Symposium on Digital Production*. 2016, pp. 13–17.

George Borshukov et al. “Universal capture-image-based facial animation for" The Matrix Reloaded"”. In: *ACM Siggraph 2005 Courses*. 2005, 16–es.

Multi-View and Motion Capture in Movies

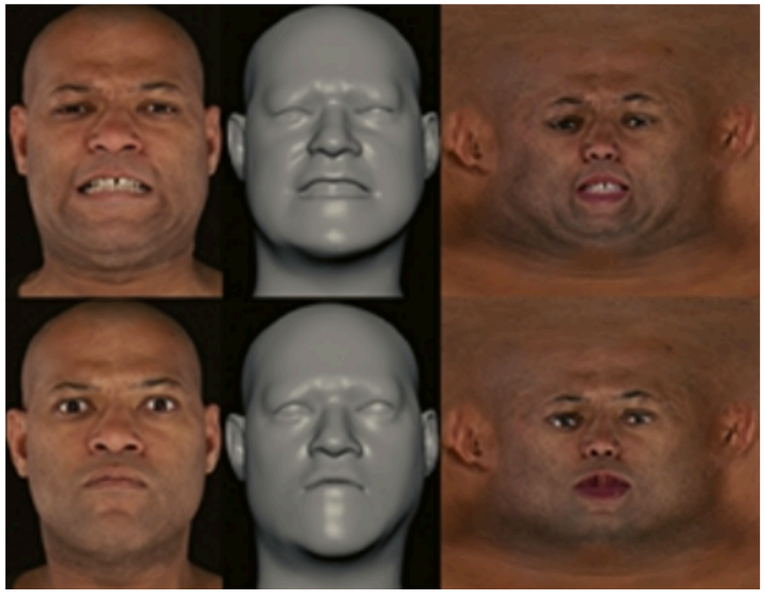


- Accurate and nuanced reconstructions
- Augment the range of artistic work

Yeongho Seol, Wan-Chun Ma, and John P Lewis. “Creating an actor-specific facial rig from performance capture”. In: *Proceedings of the 2016 Symposium on Digital Production*. 2016, pp. 13–17.

George Borshukov et al. “Universal capture-image-based facial animation for" The Matrix Reloaded"”. In: *ACM Siggraph 2005 Courses*. 2005, 16–es.

Multi-View and Motion Capture in Movies



- Accurate and nuanced reconstructions
- Augment the range of artistic work

- Expensive Technology
- Tailored to a small set of individuals
- Extensive manual refinements

Yeongho Seol, Wan-Chun Ma, and John P Lewis. “Creating an actor-specific facial rig from performance capture”. In: *Proceedings of the 2016 Symposium on Digital Production*. 2016, pp. 13–17.

George Borshukov et al. “Universal capture-image-based facial animation for" The Matrix Reloaded"”. In: *ACM Siggraph 2005 Courses*. 2005, 16–es.

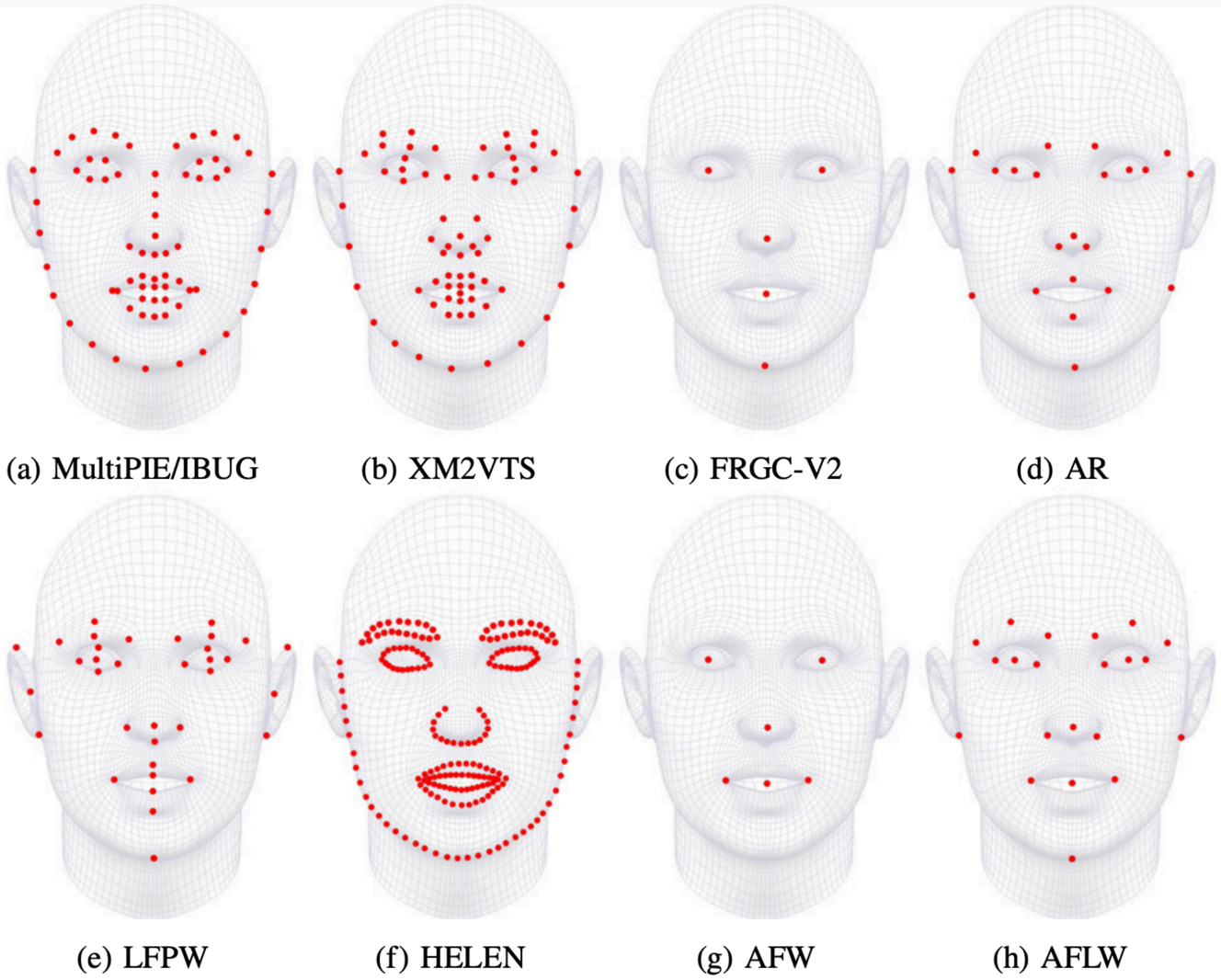
Monocular Images: A Challenging Convenient Alternative

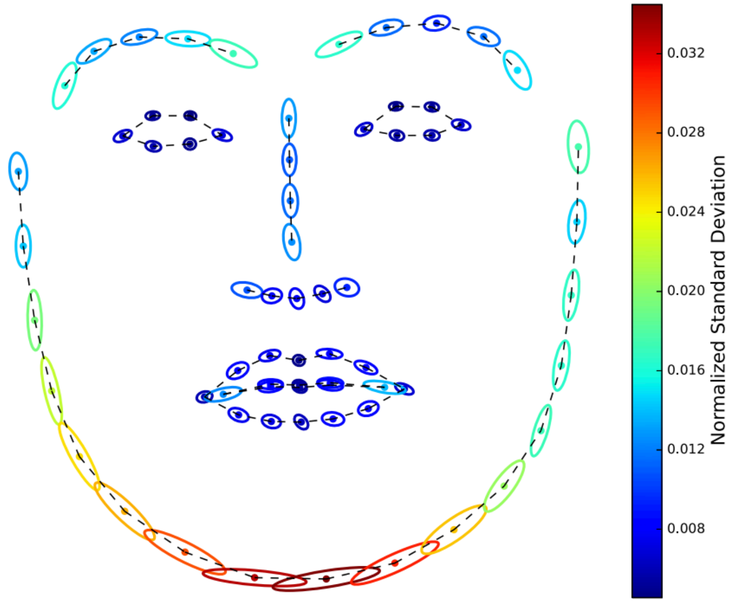
Monocular Images: A Challenging Convenient Alternative




- Occlusions
- Perspective ambiguities
- Appearance and light variations

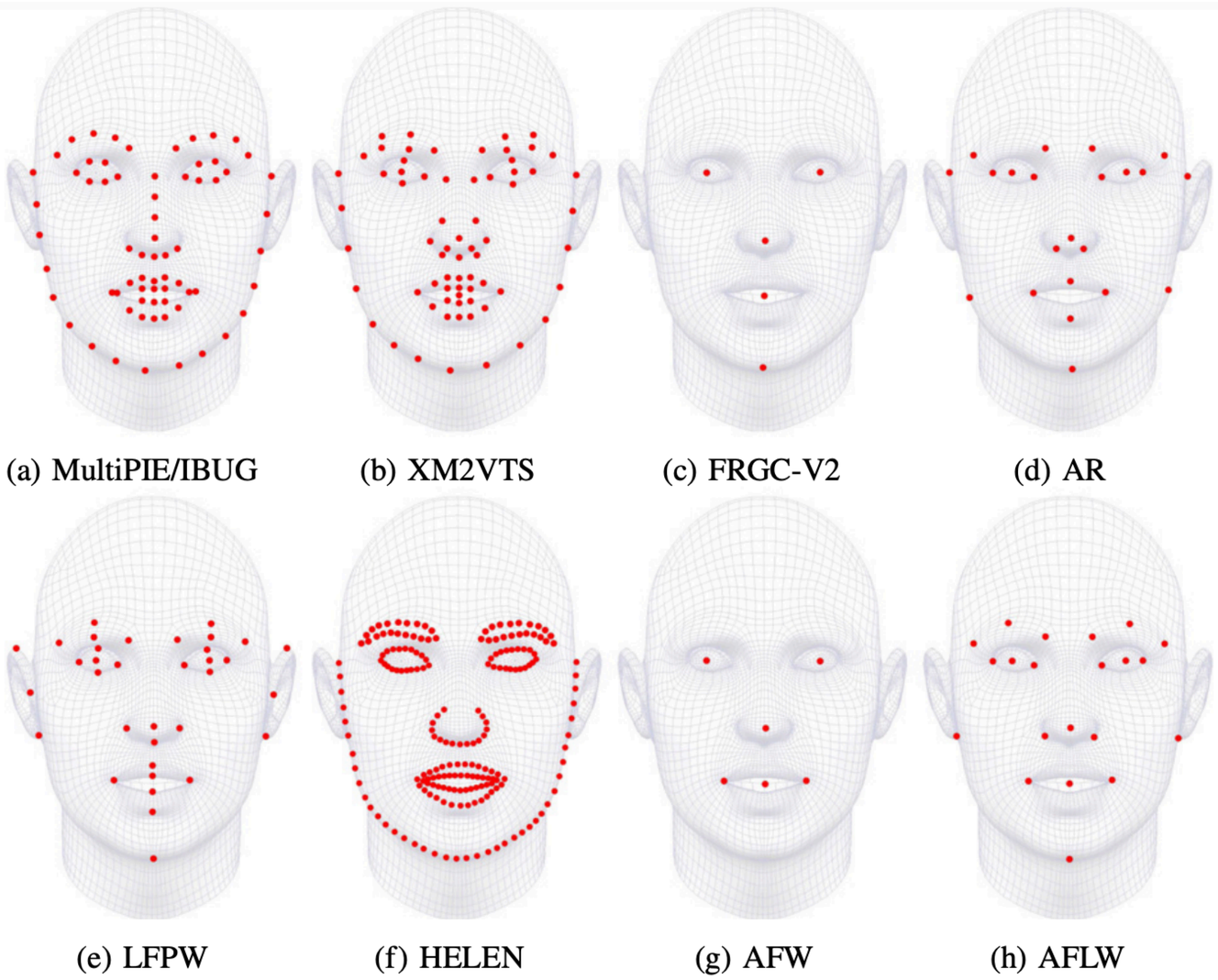
Face Representation 2D: Landmarks

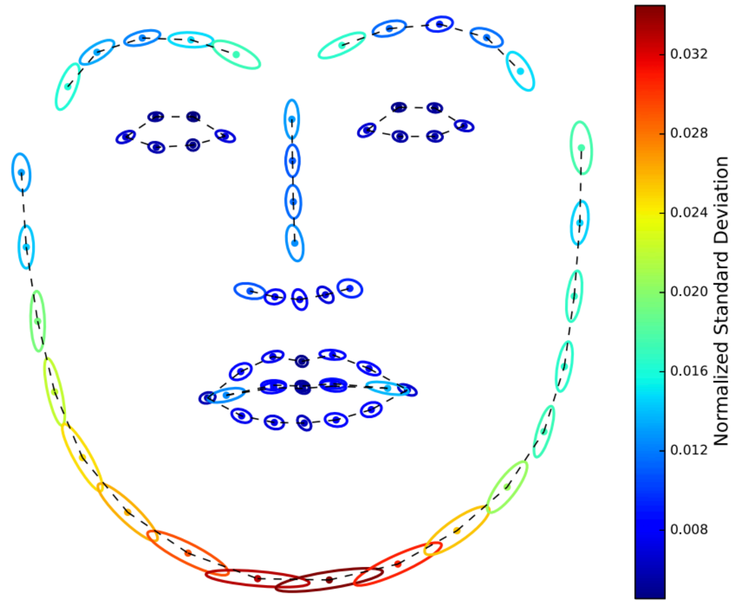




Christos Sagonas et al. “300 faces in-the-wild challenge: The first facial landmark localization challenge”. In: *Proceedings of the IEEE international conference on computer vision workshops*. 2013, pp. 397–403.

Face Representation 2D: Landmarks



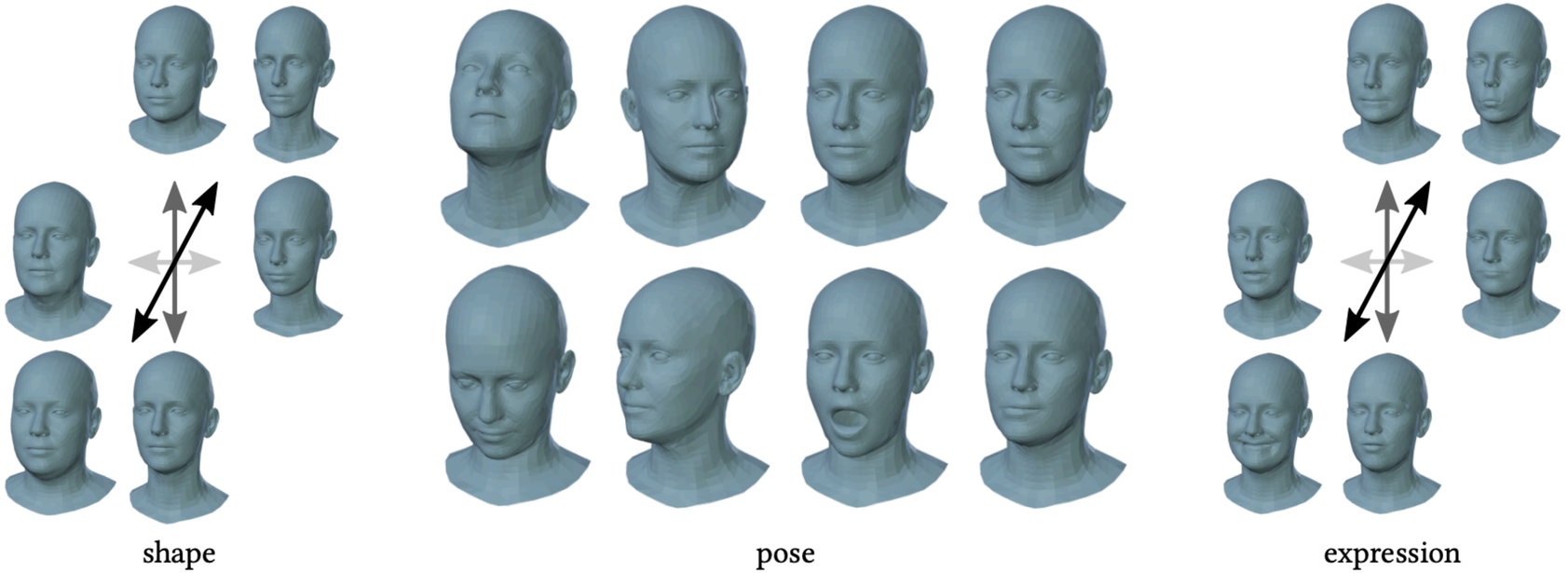


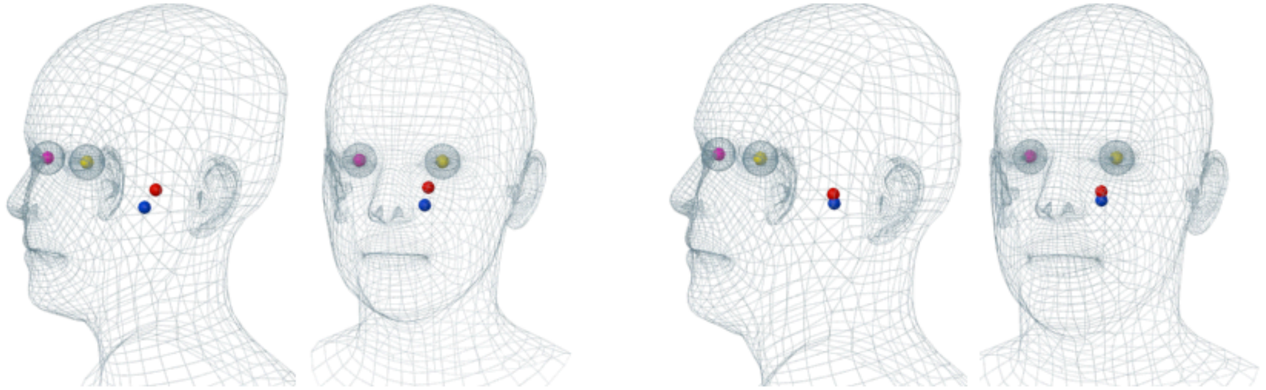
Sparse

Christos Sagonas et al. “300 faces in-the-wild challenge: The first facial landmark localization challenge”. In: *Proceedings of the IEEE international conference on computer vision workshops*. 2013, pp. 397–403.

Face Representation 3D: Parametric Head Model

FLAME

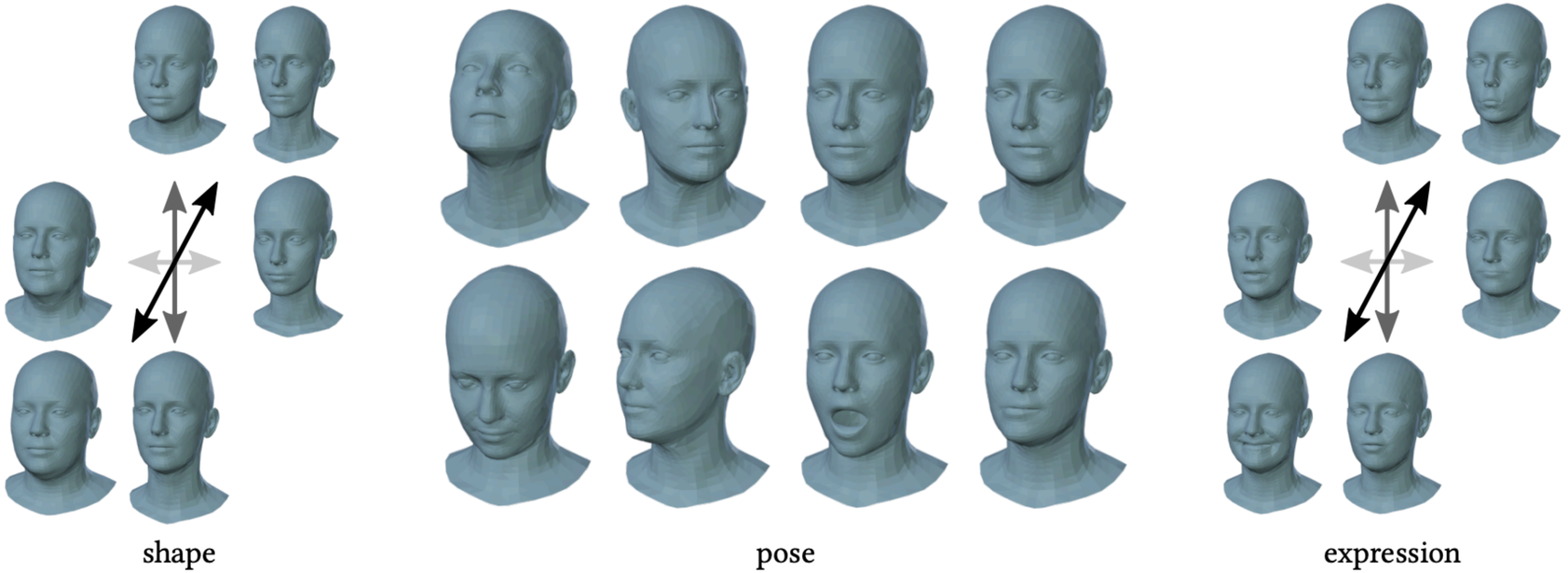


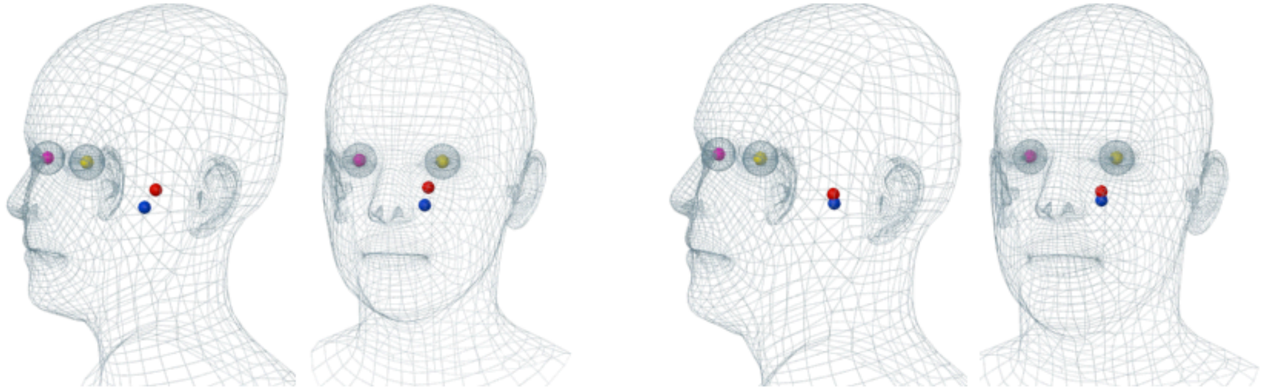


Tianye Li et al. “Learning a model of facial shape and expression from 4D scans”. In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), 194:1–194:17. URL: https://doi.org/10.1145/3130800.3130813.

Face Representation 3D: Parametric Head Model

FLAME

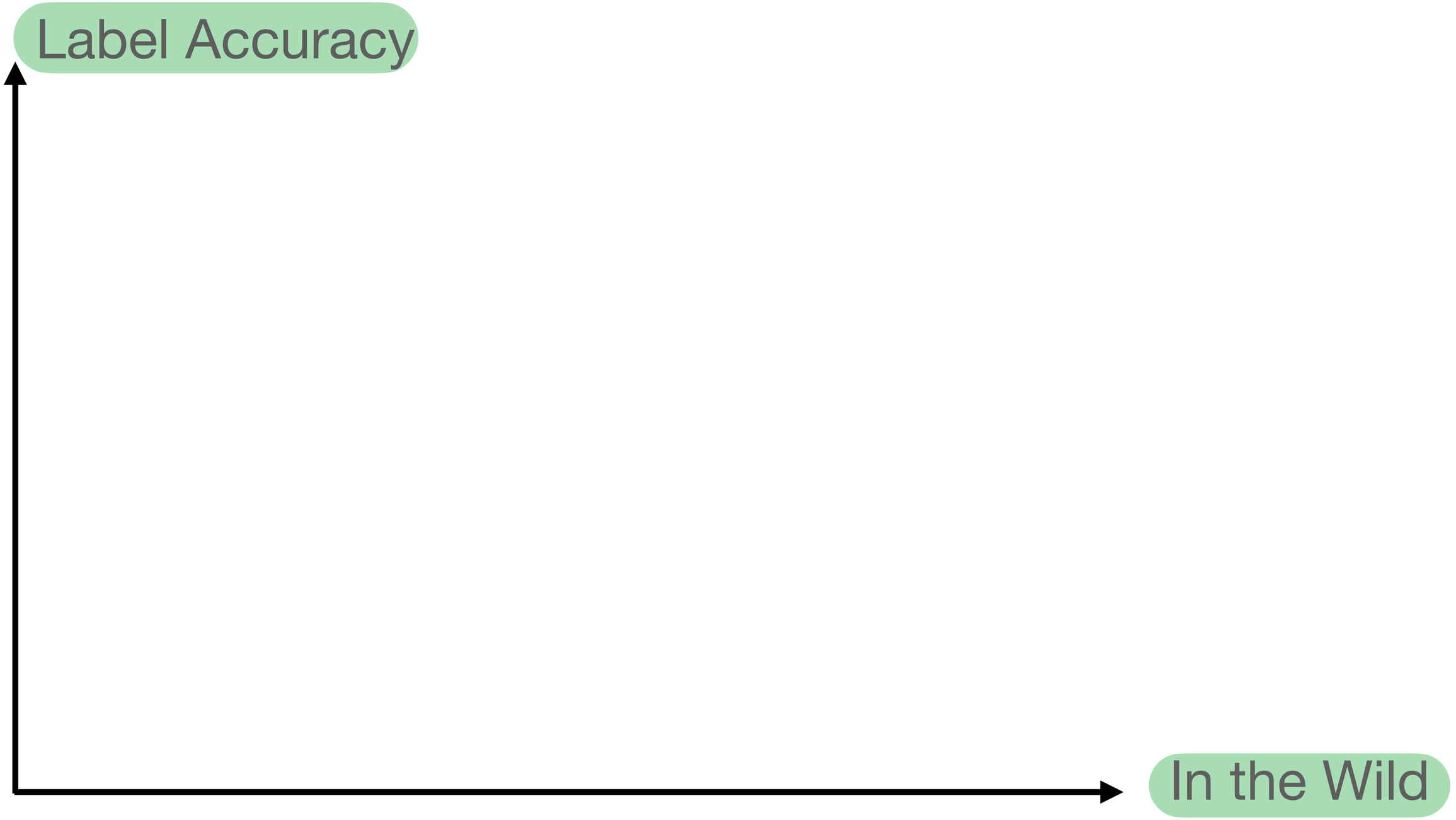




- Compact representation
- Adds the 3D prior

Tianye Li et al. “Learning a model of facial shape and expression from 4D scans”. In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), 194:1–194:17. URL: https://doi.org/10.1145/3130800.3130813.

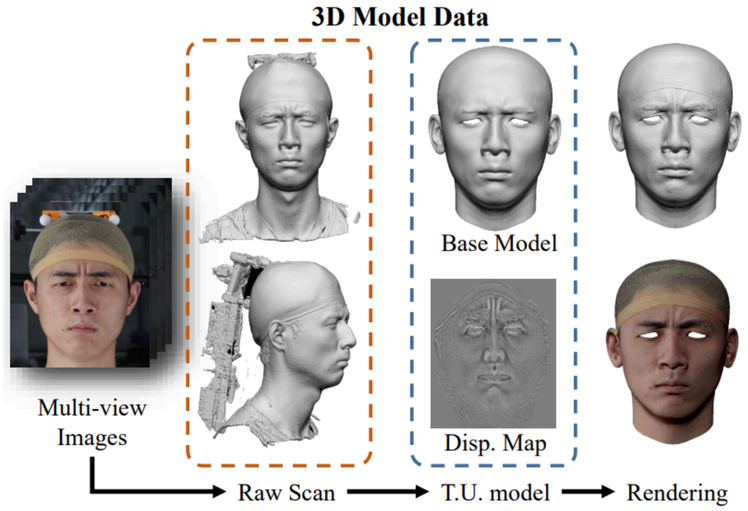
Datasets



Label Accuracy

In the Wild

Multi-view Images:
FaceScape

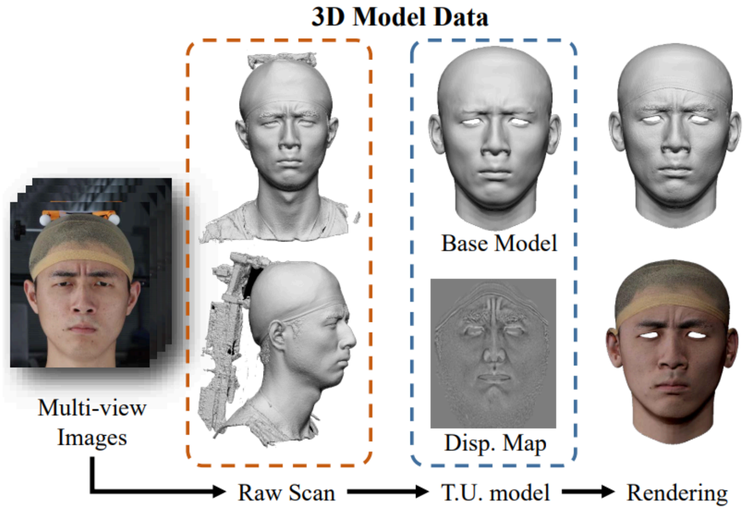


Haotian Yang et al. "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, pp. 601–610.

Label Accuracy

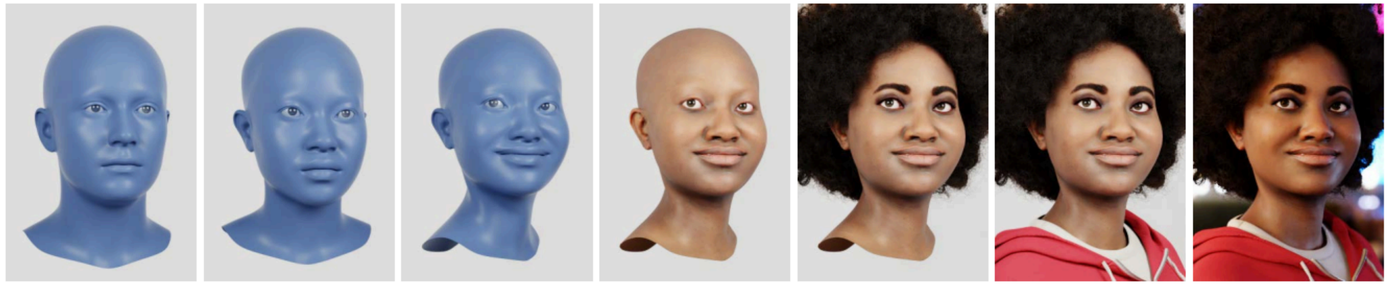
In the Wild

- Multi-view Images: **FaceScape**



Haotian Yang et al. "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, pp. 601–610.

- Synthetic Images: **FaceSynthetics**

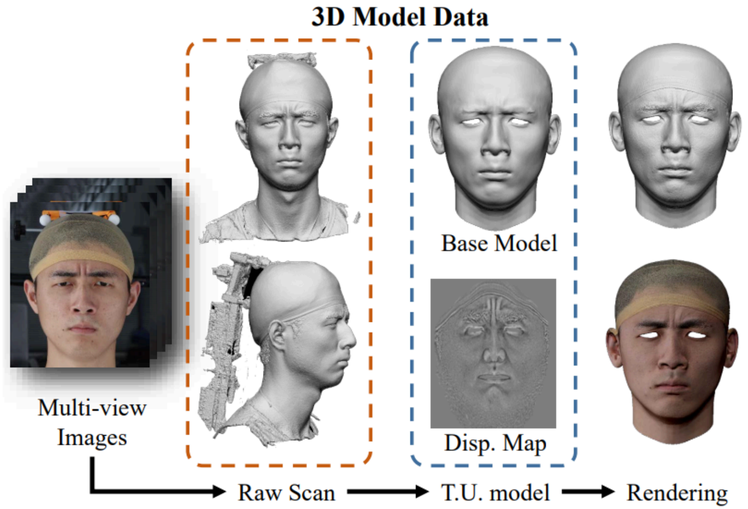


Erroll Wood et al. "Fake it till you make it: face analysis in the wild using synthetic data alone". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3681–3691.

Label Accuracy

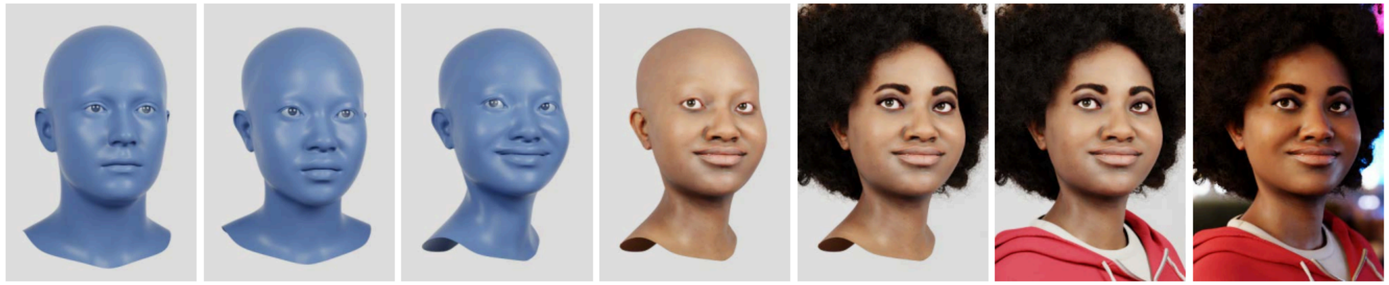
In the Wild

Multi-view Images: **FaceScape**



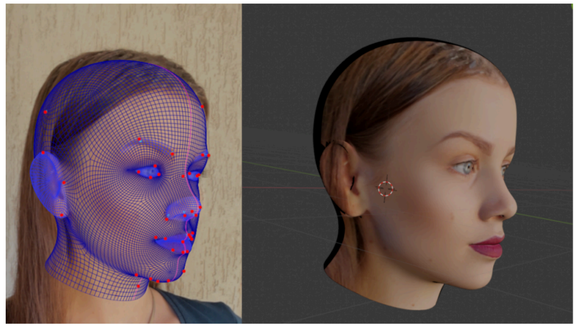
Haotian Yang et al. "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, pp. 601–610.

Synthetic Images: **FaceSynthetics**



Erroll Wood et al. "Fake it till you make it: face analysis in the wild using synthetic data alone". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3681–3691.

Fit a Model: **DAD-3DHead**

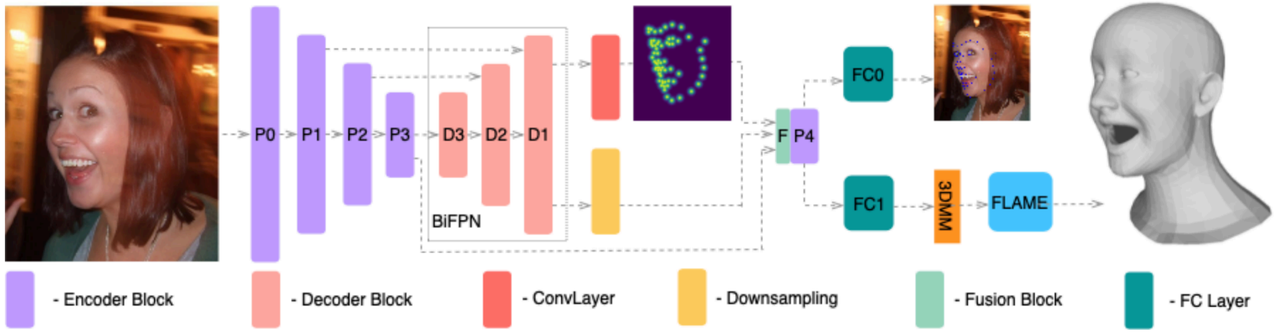


Tetiana Martyniuk et al. "DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20942–20952.

Related Works

Noisy labels	In the wild images
Sparse Landmarks	
FLAME prediction	
Open source	

DAD-3DNet



Tetiana Martyniuk et al. “DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20942–20952.

Related Works

Noisy labels

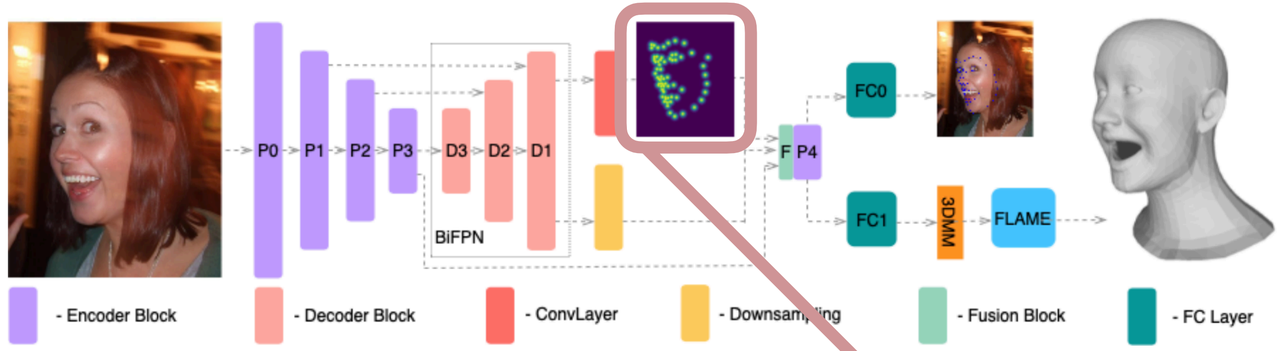
In the wild images

Sparse Landmarks

FLAME prediction

Open source

DAD-3DNet



Heatmap Regression

Spatial complexity limits number of landmarks

Tetiana Martyniuk et al. “DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20942–20952.

Related Works

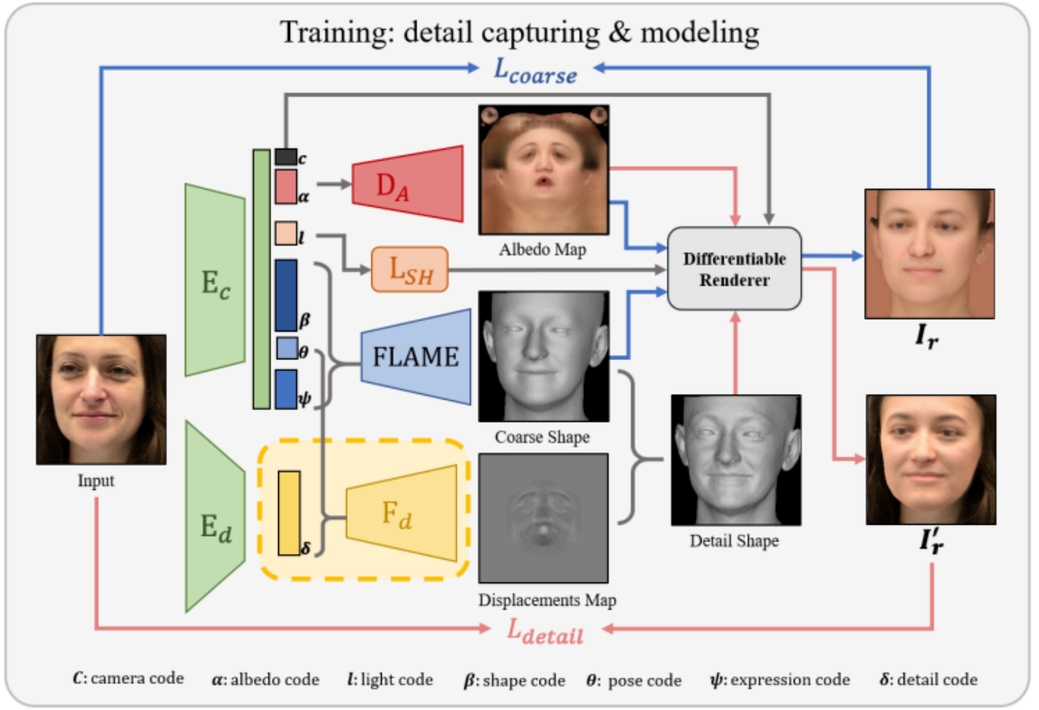
Analysis by Synthesis

Sparse Landmarks

FLAME prediction

Open source

DECA



Yao Feng et al. "Learning an Animatable Detailed 3D Face Model from In-the-Wild Images". In: *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH* 40.4 (Aug. 2021), 88:1–88:13.

Related Works

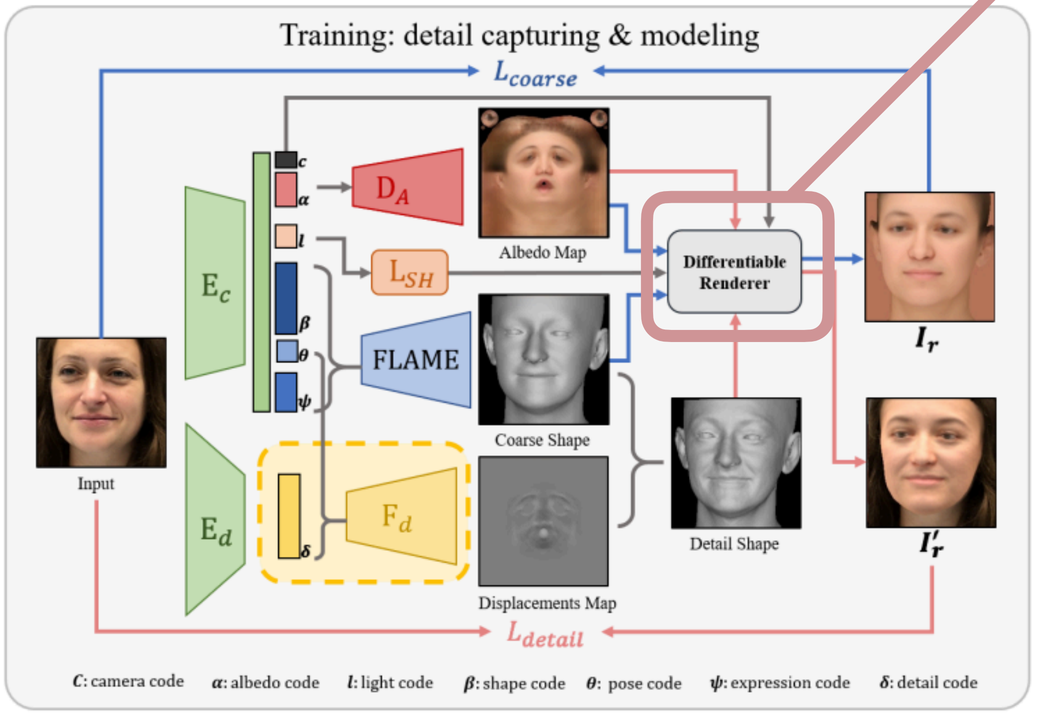
Analysis by Synthesis

Sparse Landmarks

FLAME prediction

Open source

DECA



Differentiable Renderer

Need approximations for illumination, facial appearance and rendering

Yao Feng et al. "Learning an Animatable Detailed 3D Face Model from In-the-Wild Images". In: *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH* 40.4 (Aug. 2021), 88:1–88:13.

Related Works

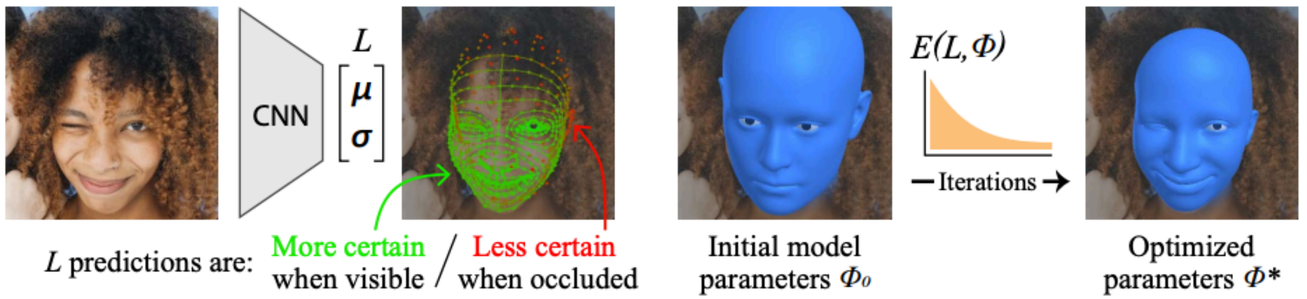
Synthetic images

Distribution over Dense Landmarks

FLAME fitting

Closed source

Wood et al.



Erroll Wood et al. “3d face reconstruction with dense landmarks”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 160–177.

Related Works

Great Complexity

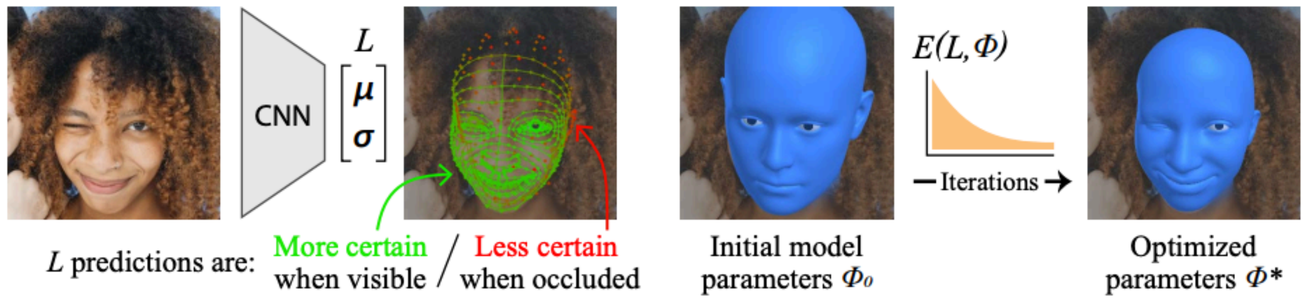
Synthetic images

Distribution over Dense Landmarks

FLAME fitting

Closed source

Wood et al.



Erroll Wood et al. “3d face reconstruction with dense landmarks”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 160–177.

Related Works

Great Complexity

Gaussian Prior

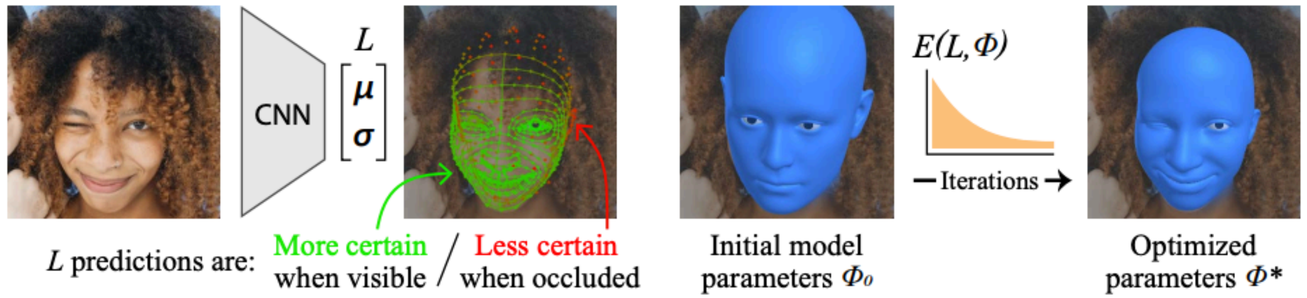
Synthetic images

Distribution over Dense Landmarks

FLAME fitting

Closed source

Wood et al.



Erroll Wood et al. “3d face reconstruction with dense landmarks”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 160–177.

Related Works

Great Complexity

Gaussian Prior

Analytic Jacobians

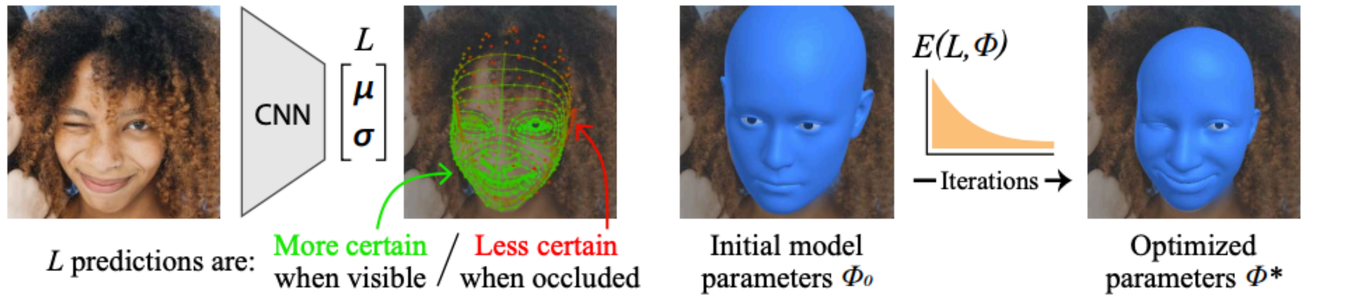
Synthetic images

Distribution over Dense Landmarks

FLAME fitting

Closed source

Wood et al.



Erroll Wood et al. “3d face reconstruction with dense landmarks”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 160–177.

Related Works

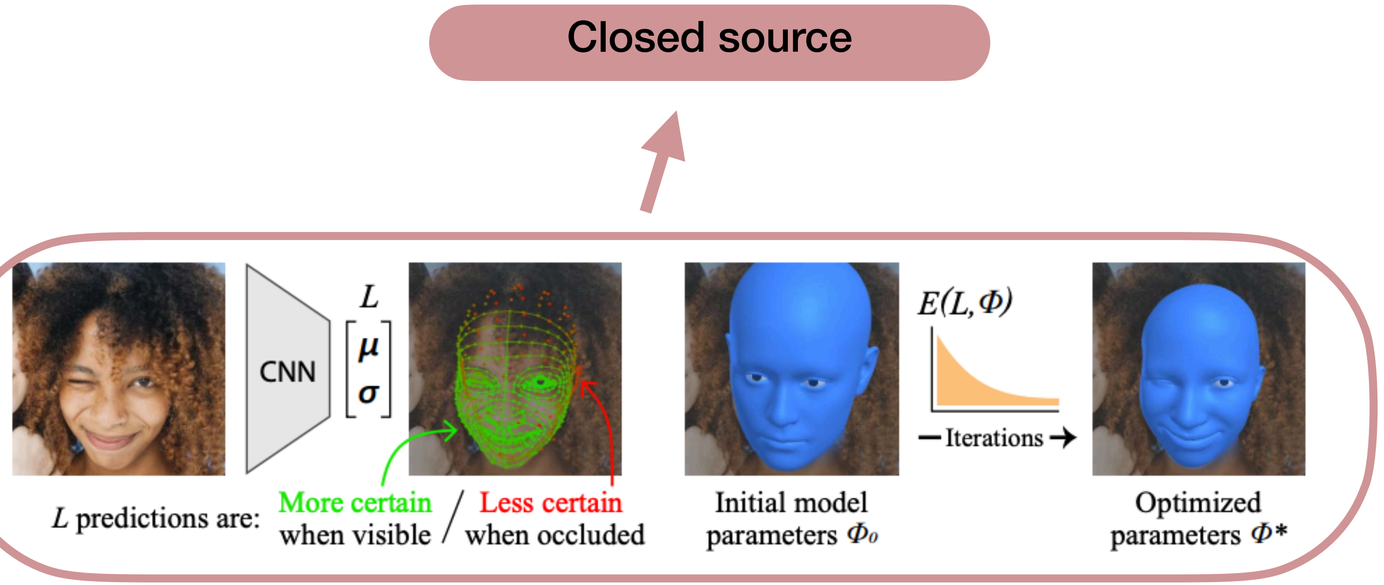
Synthetic images

Distribution over Dense Landmarks

FLAME fitting

Closed source

Wood et al.



Erroll Wood et al. “3d face reconstruction with dense landmarks”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 160–177.

Noisy labels

In the wild images

Sparse Landmarks

FLAME prediction

Open source

Analysis by Synthesis

Sparse Landmarks

FLAME prediction

Open source

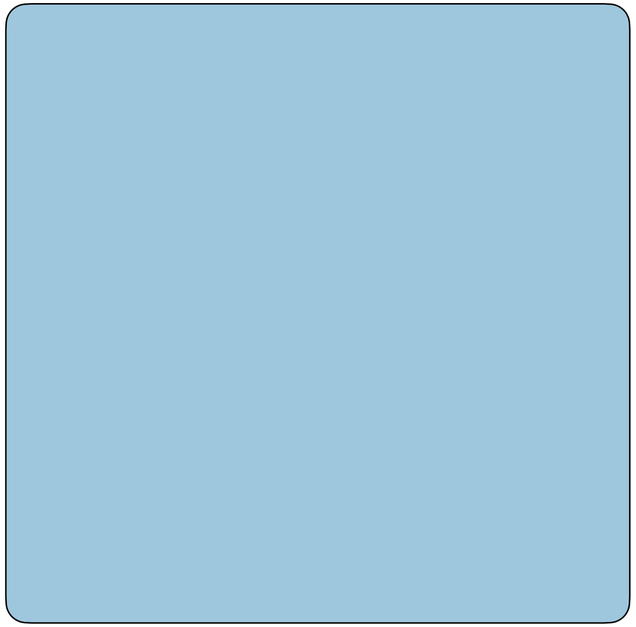
Synthetic images

Distribution over Dense Landmarks

FLAME fitting

Closed source

Our Method



Noisy labels

In the wild images

Sparse Landmarks

FLAME prediction

Open source

Analysis by Synthesis

Sparse Landmarks

FLAME prediction

Open source

Synthetic images

Distribution over Dense Landmarks

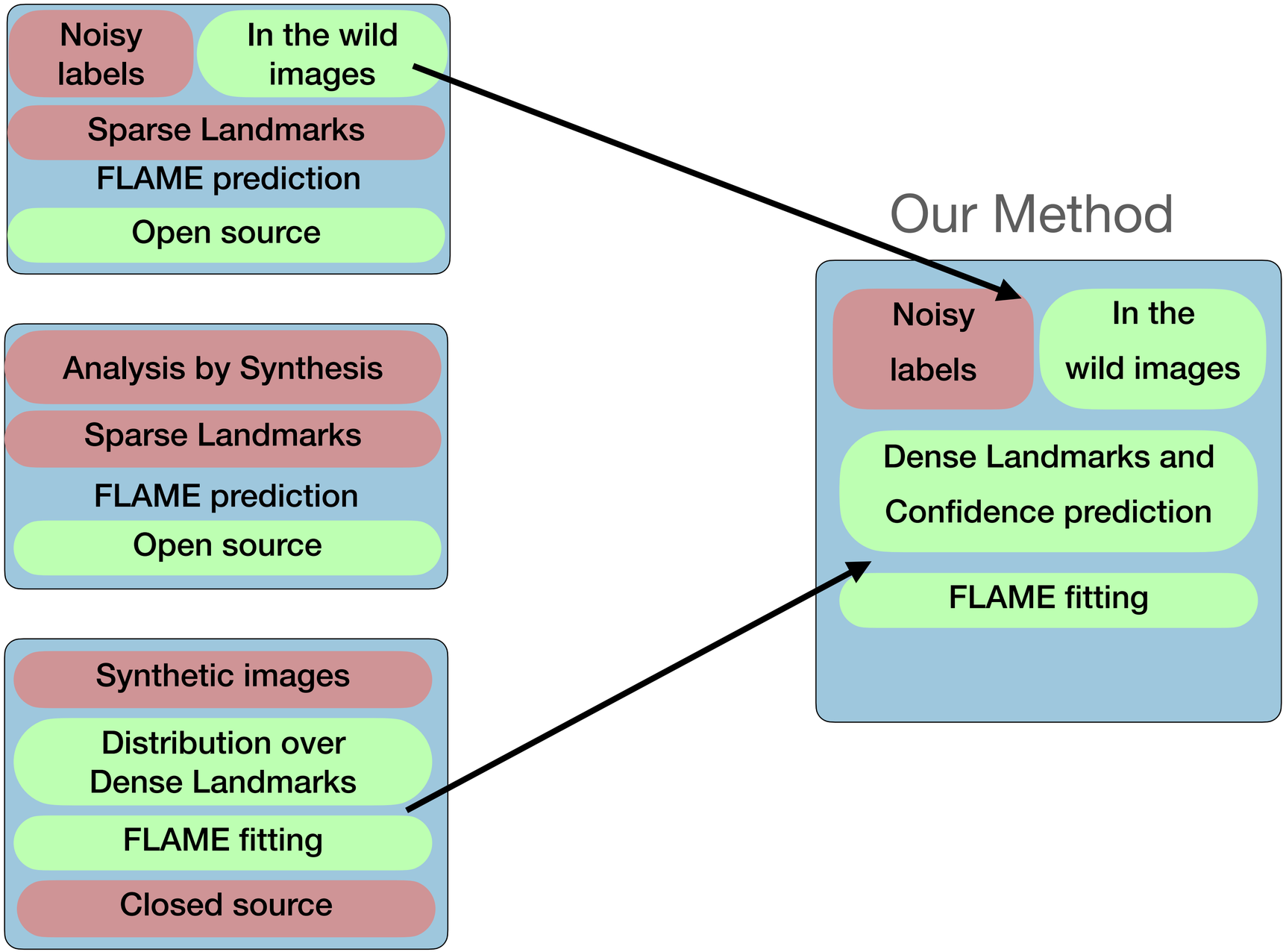
FLAME fitting

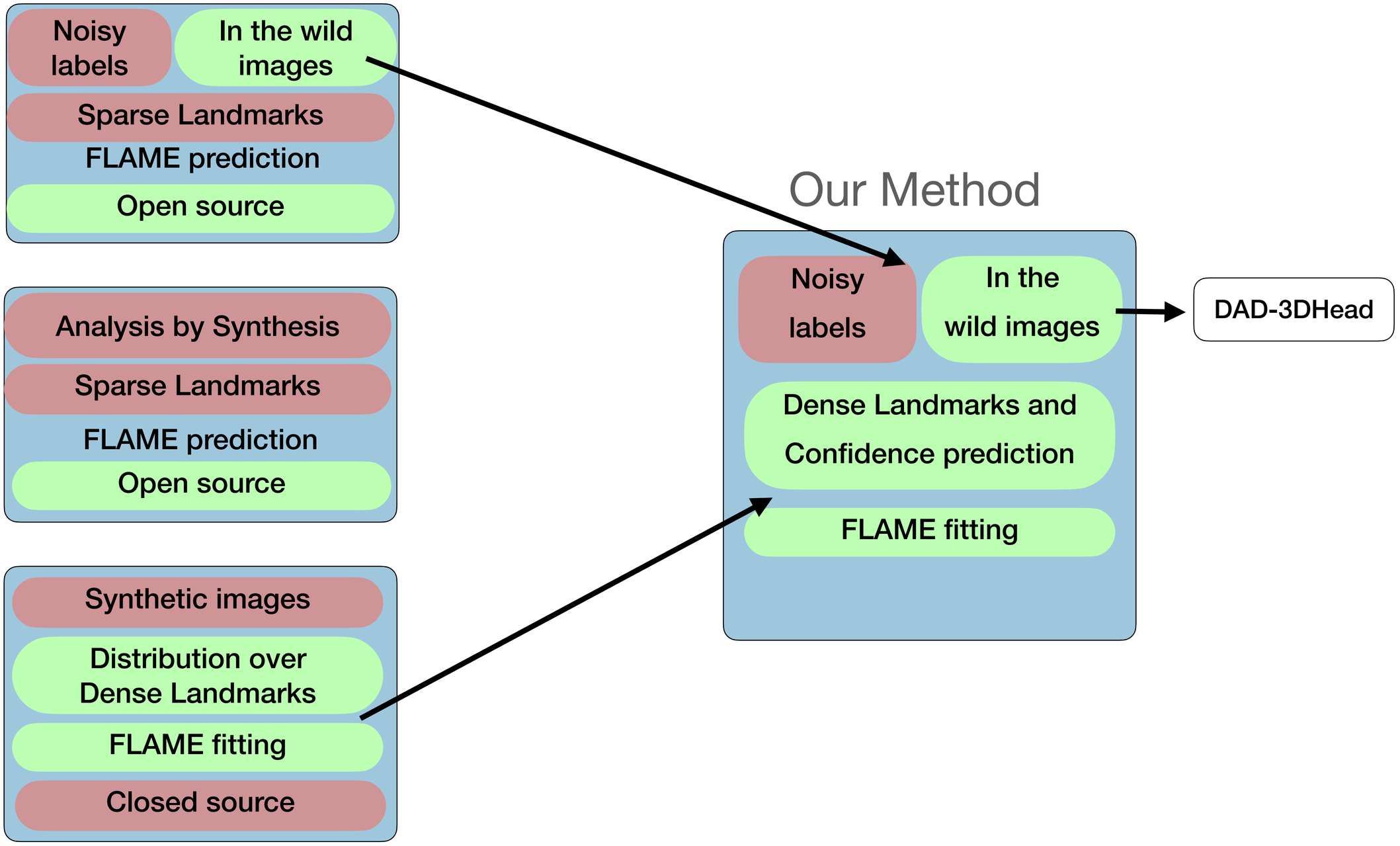
Closed source

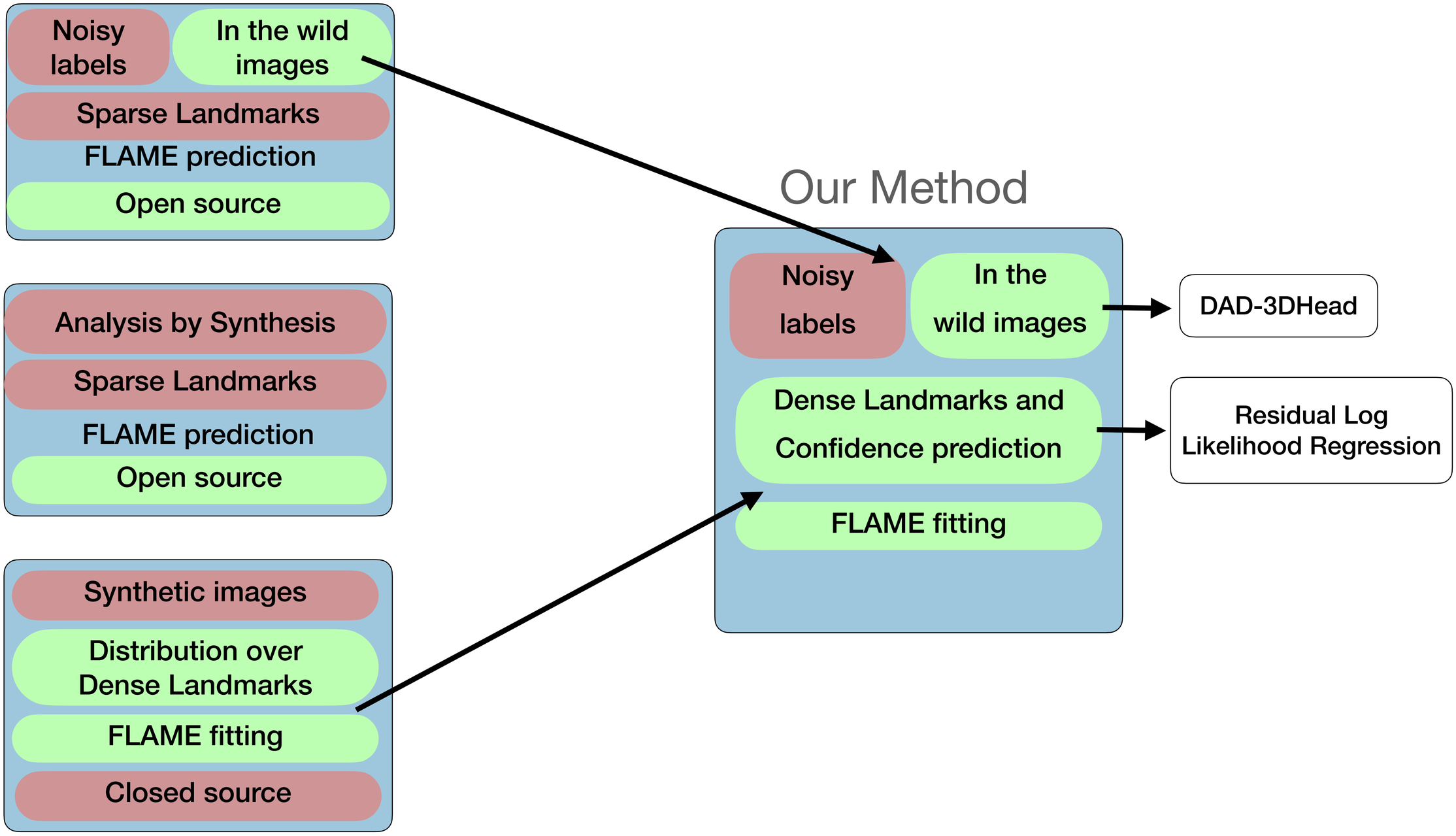
Our Method

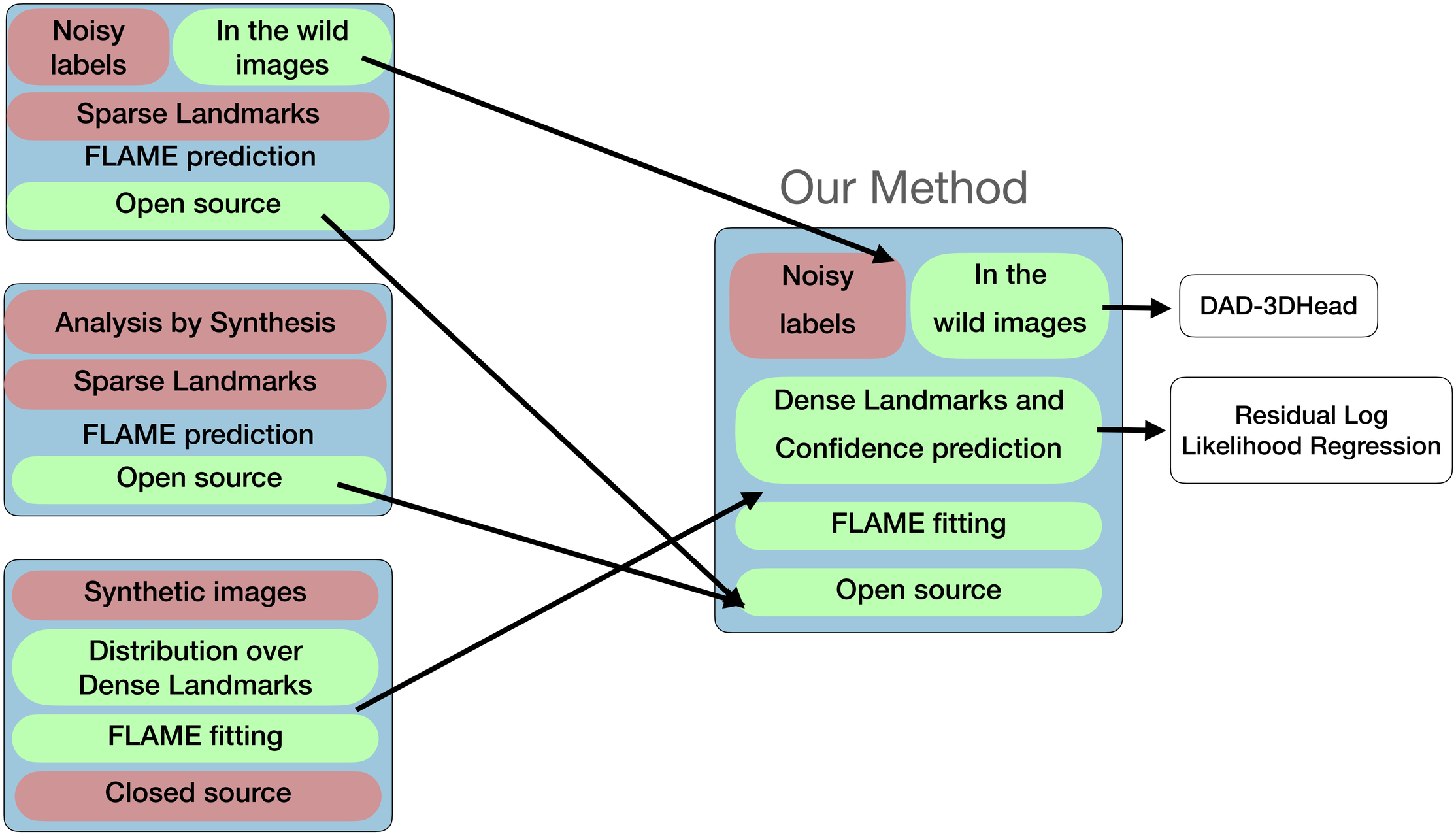
Dense Landmarks and Confidence prediction

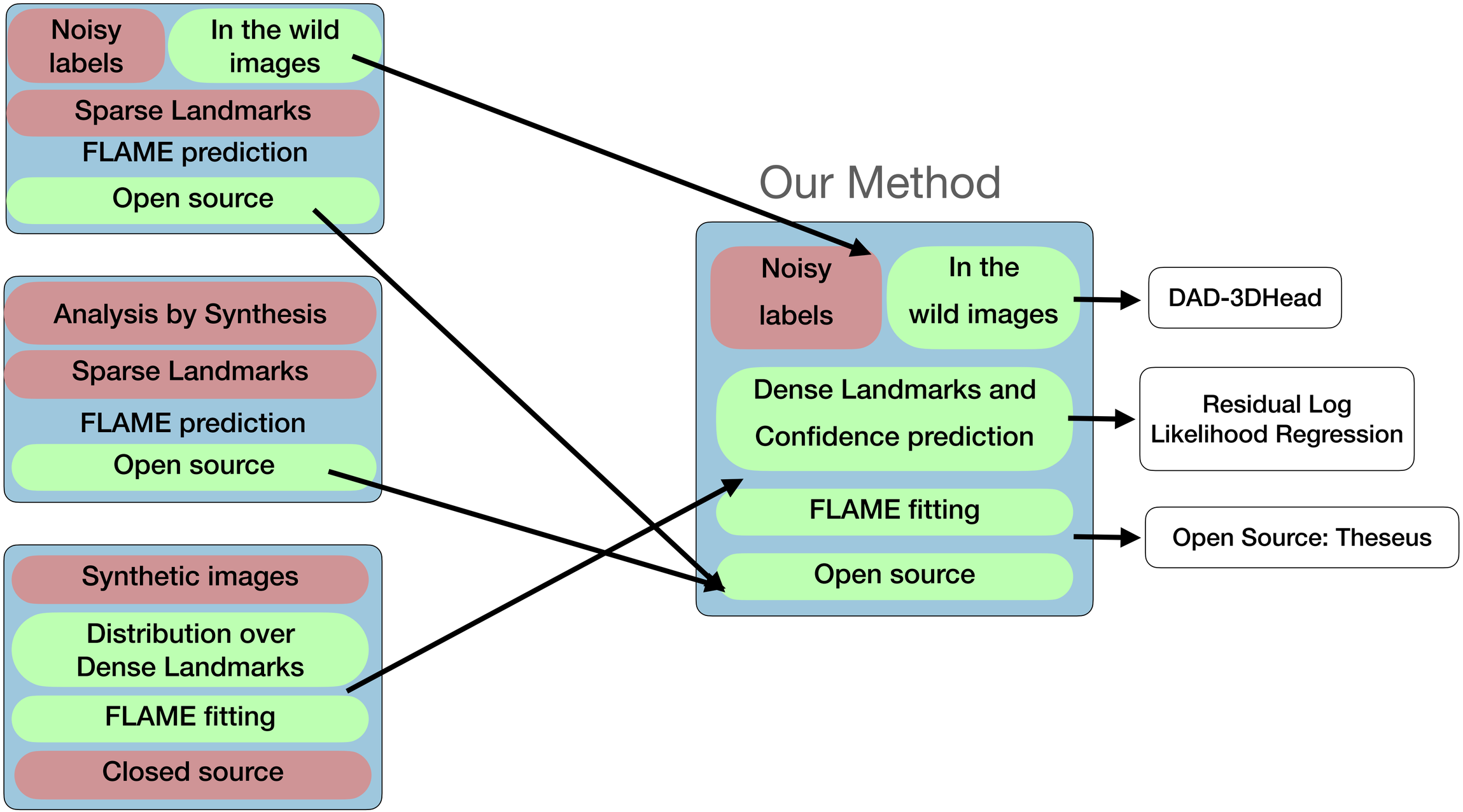
FLAME fitting


Noisy labels
In the wild images
Sparse Landmarks
FLAME prediction
Open source
Analysis by Synthesis
Sparse Landmarks
FLAME prediction
Open source
Synthetic images
Distribution over Dense Landmarks
FLAME fitting
Closed source
Our Method
Noisy labels
In the wild images
Dense Landmarks and Confidence prediction
FLAME fitting

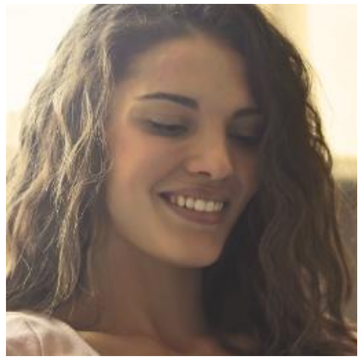

Noisy labels
In the wild images
Sparse Landmarks
FLAME prediction
Open source
Analysis by Synthesis
Sparse Landmarks
FLAME prediction
Open source
Synthetic images
Distribution over Dense Landmarks
FLAME fitting
Closed source
Our Method
Noisy labels
In the wild images
Dense Landmarks and Confidence prediction
FLAME fitting
DAD-3DHead


Noisy labels
In the wild images
Sparse Landmarks
FLAME prediction
Open source
Analysis by Synthesis
Sparse Landmarks
FLAME prediction
Open source
Synthetic images
Distribution over Dense Landmarks
FLAME fitting
Closed source
Our Method
Noisy labels
In the wild images
Dense Landmarks and Confidence prediction
FLAME fitting
DAD-3DHead
Residual Log Likelihood Regression


Noisy labels
In the wild images
Sparse Landmarks
FLAME prediction
Open source
Analysis by Synthesis
Sparse Landmarks
FLAME prediction
Open source
Synthetic images
Distribution over Dense Landmarks
FLAME fitting
Closed source
Our Method
Noisy labels
In the wild images
Dense Landmarks and Confidence prediction
FLAME fitting
Open source
DAD-3DHead
Residual Log Likelihood Regression


Noisy labels
In the wild images
Sparse Landmarks
FLAME prediction
Open source
Analysis by Synthesis
Sparse Landmarks
FLAME prediction
Open source
Synthetic images
Distribution over Dense Landmarks
FLAME fitting
Closed source
Our Method
Noisy labels
In the wild images
Dense Landmarks and Confidence prediction
FLAME fitting
Open source
DAD-3DHead
Residual Log Likelihood Regression
Open Source: Theseus

Method



Noisy labels

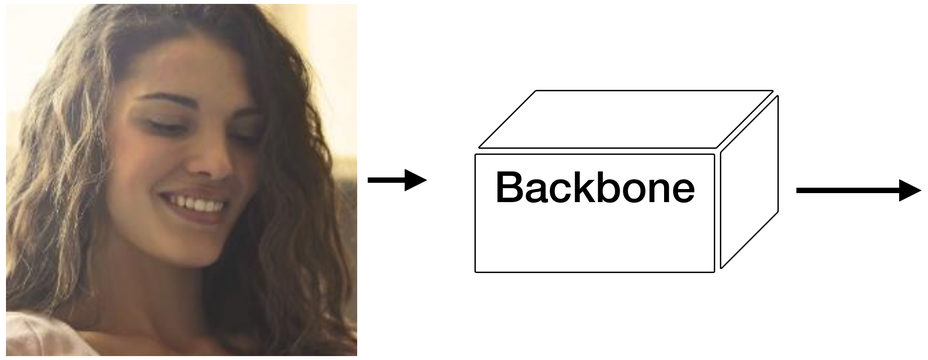
In the wild images

Dense Landmarks and Confidence prediction

FLAME fitting

Open source

Method



Noisy labels

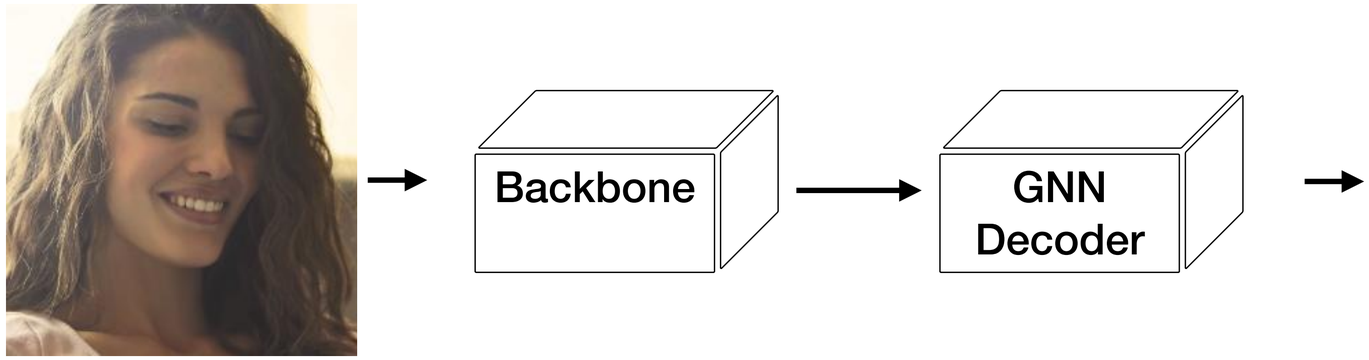
In the wild images

Dense Landmarks and Confidence prediction

FLAME fitting

Open source

Method



Noisy labels

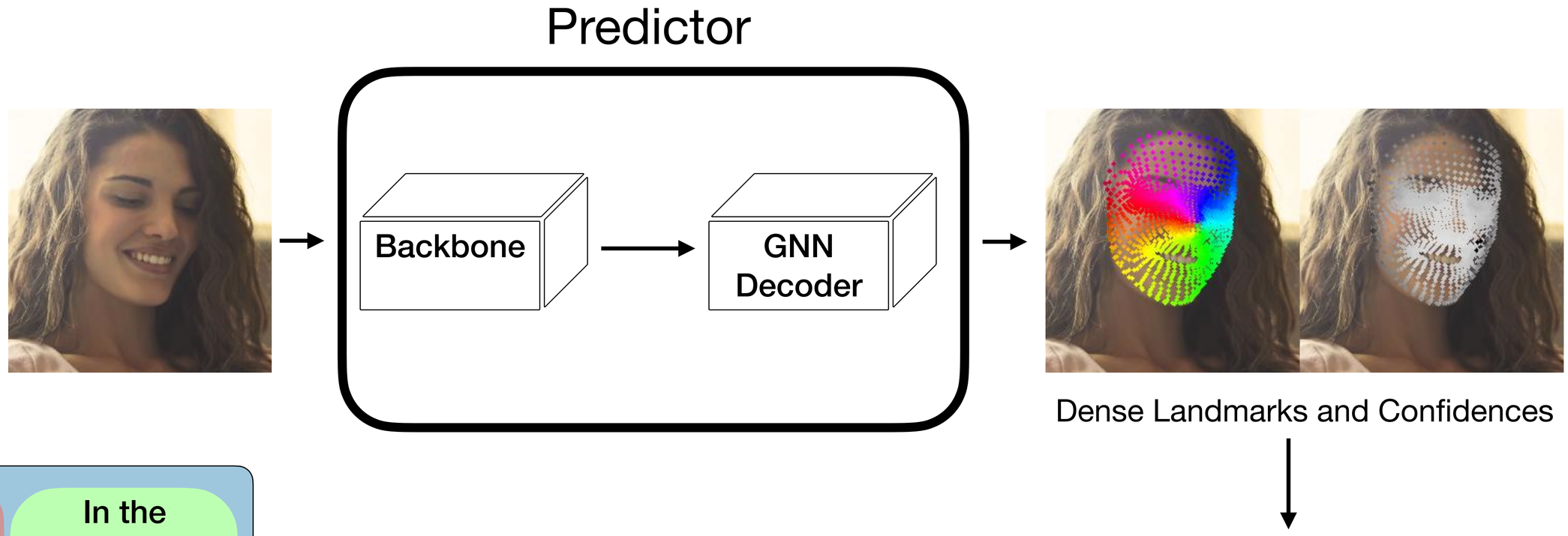
In the wild images

Dense Landmarks and Confidence prediction

FLAME fitting

Open source

Method



Noisy labels

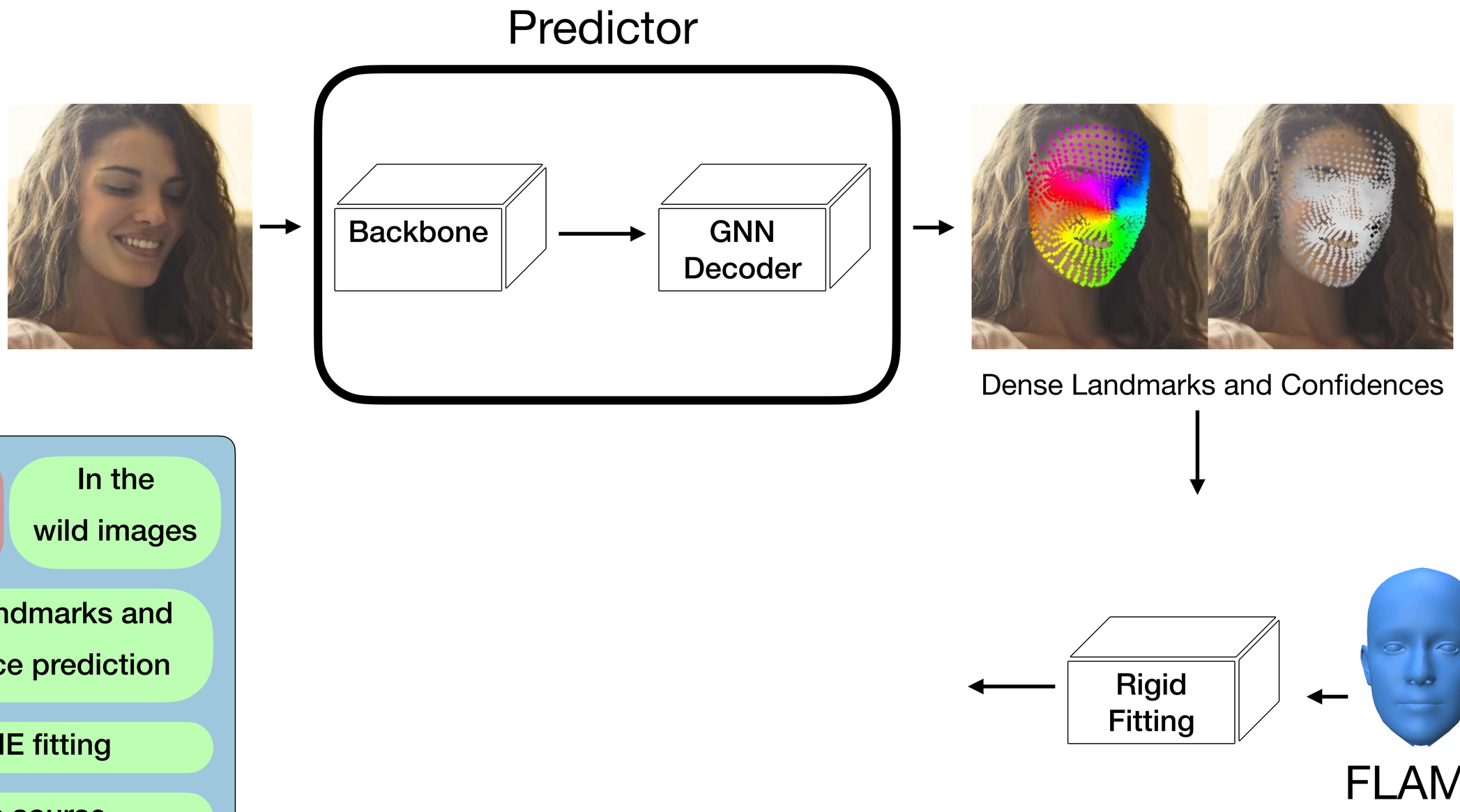
In the wild images

Dense Landmarks and Confidence prediction

FLAME fitting

Open source

Method



Noisy labels

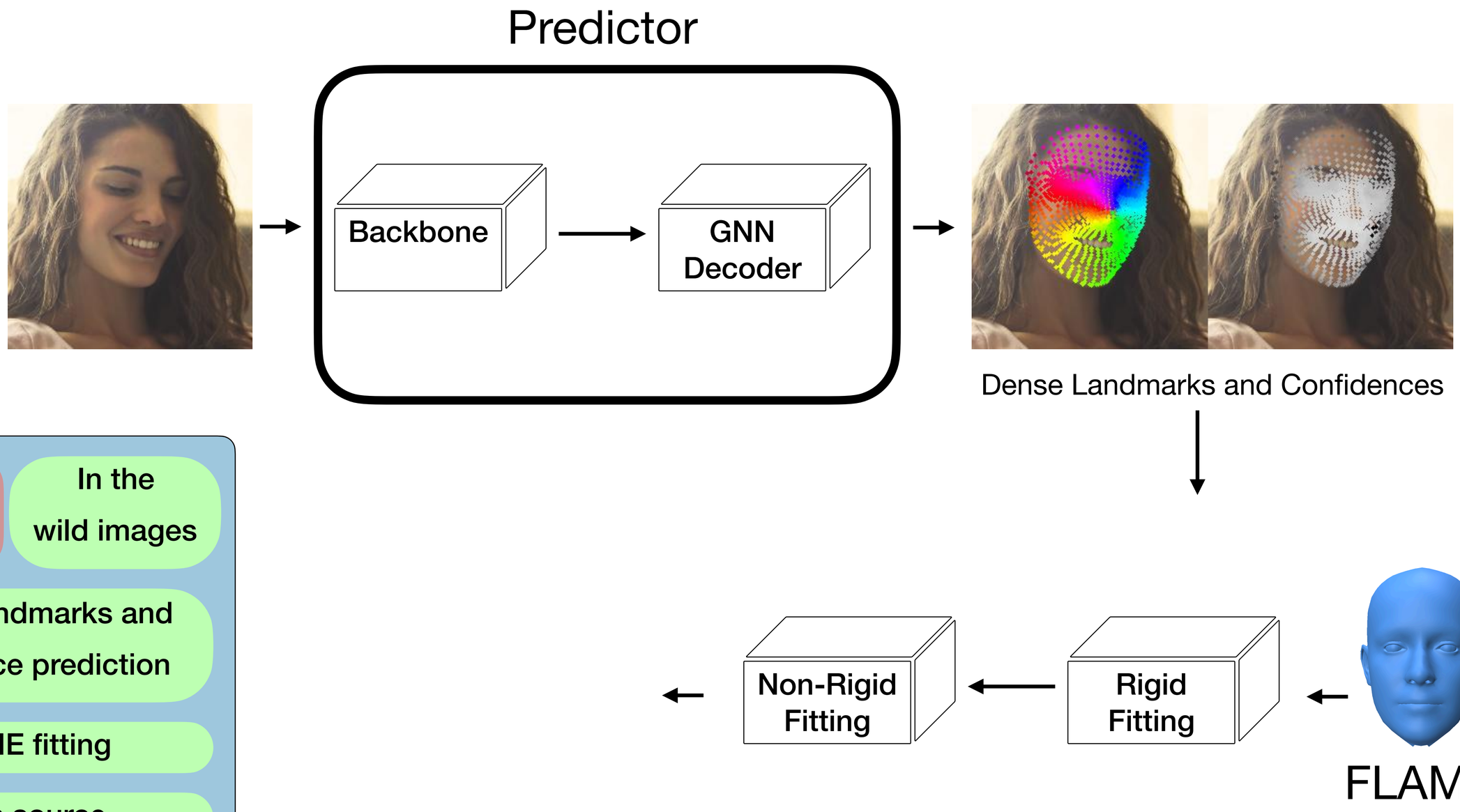
In the wild images

Dense Landmarks and Confidence prediction

FLAME fitting

Open source

Method



Noisy labels

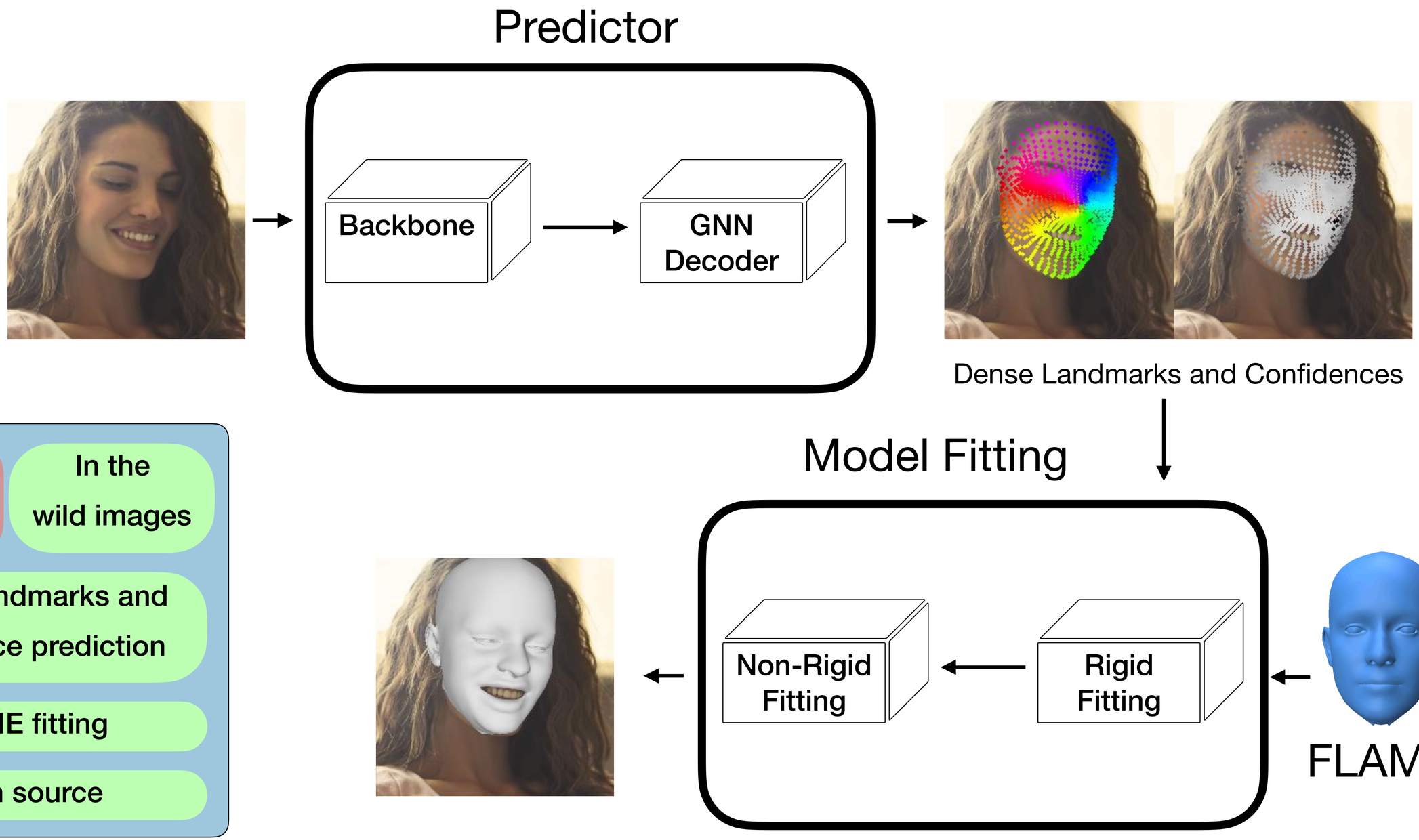
In the wild images

Dense Landmarks and Confidence prediction

FLAME fitting

Open source

Method



Noisy labels

In the wild images

Dense Landmarks and Confidence prediction

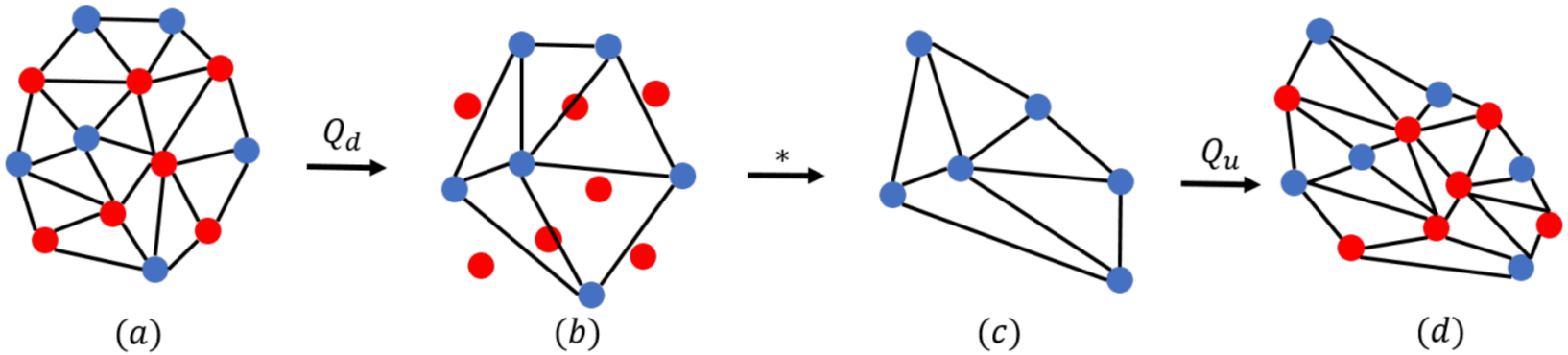
FLAME fitting

Open source

GNN Decoder: SpiralNet ++

COMA Mesh Sampling

$$U^{k-1} = Q_u^{k-1} X^{k-1}$$



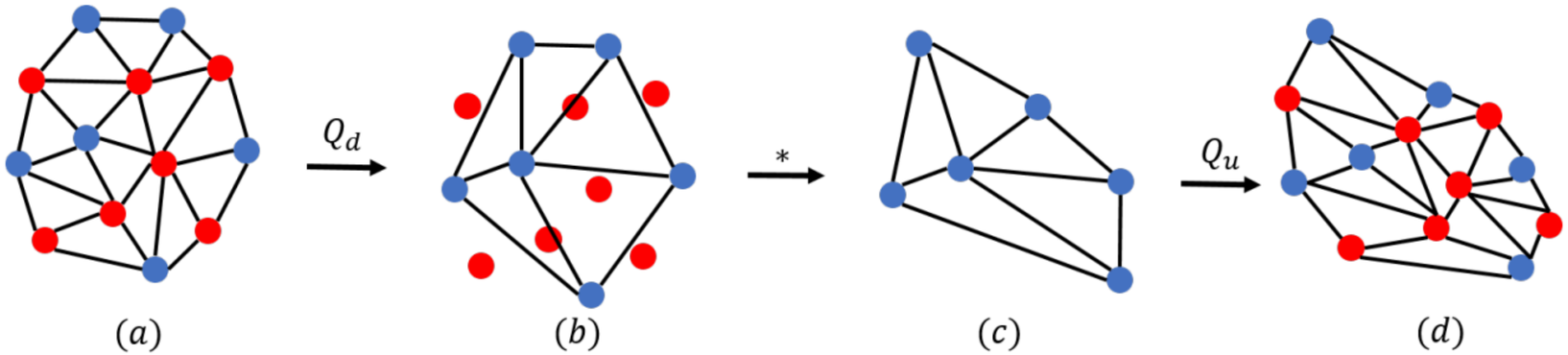
Anurag Ranjan et al. “Generating 3D faces using convolutional mesh autoencoders”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 704–720.

Shunwang Gong et al. “Spiralnet++: A fast and highly efficient mesh convolution operator”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

GNN Decoder: SpiralNet ++

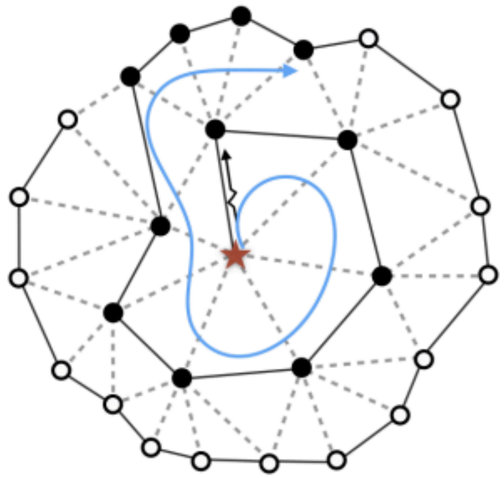
COMA Mesh Sampling

$$U^{k-1} = Q_u^{k-1} X^{k-1}$$



Spiral Convolution

$$Conv_i(U^{k-1}) = A^{k-1} \,||_{j \in S_{i,l}}\, u_j^{k-1}$$



(a) Spiral++

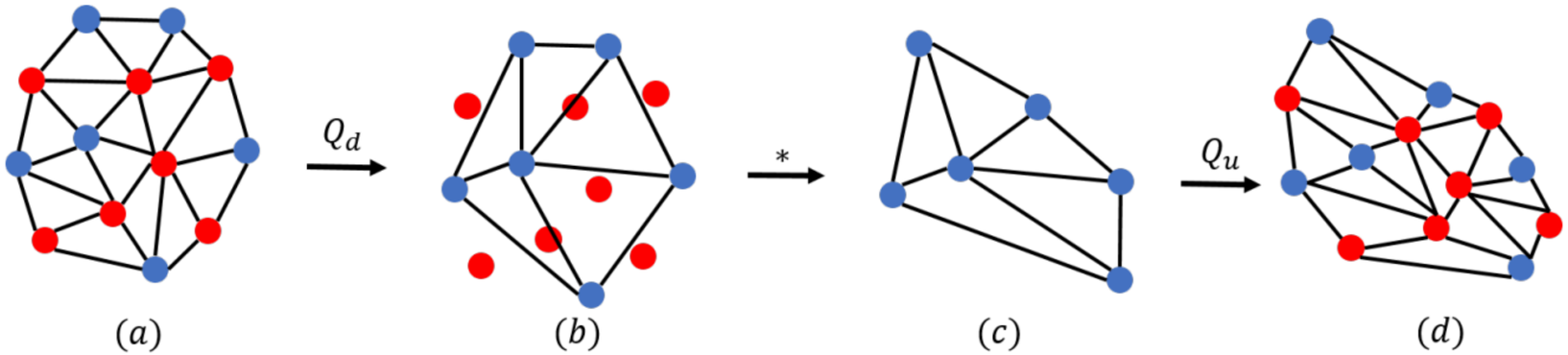
Anurag Ranjan et al. "Generating 3D faces using convolutional mesh autoencoders". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 704–720.

Shunwang Gong et al. "Spiralnet++: A fast and highly efficient mesh convolution operator". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

GNN Decoder: SpiralNet ++

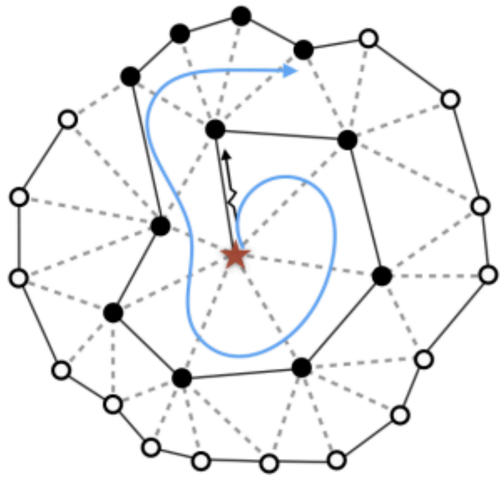
COMA Mesh Sampling

$$U^{k-1} = Q_u^{k-1} X^{k-1}$$



Spiral Convolution

$$Conv_i(U^{k-1}) = A^{k-1} \,||_{j \in S_{i,l}}\, u_j^{k-1}$$



(a) Spiral++

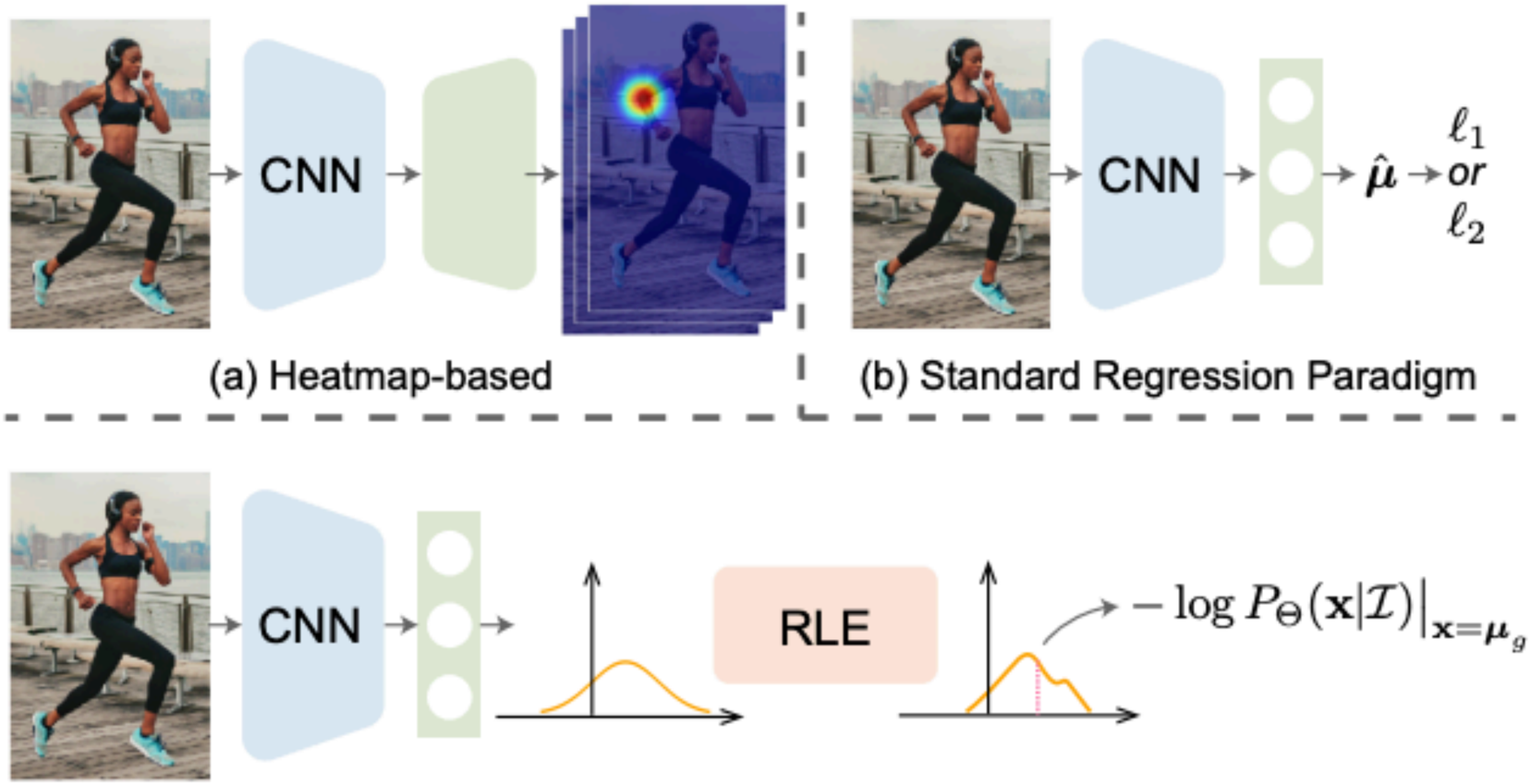
SpiralNet ++

$$x_i^k = ELU[Conv_i(U^{k-1})]$$

Anurag Ranjan et al. “Generating 3D faces using convolutional mesh autoencoders”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 704–720.

Shunwang Gong et al. “Spiralnet++: A fast and highly efficient mesh convolution operator”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.

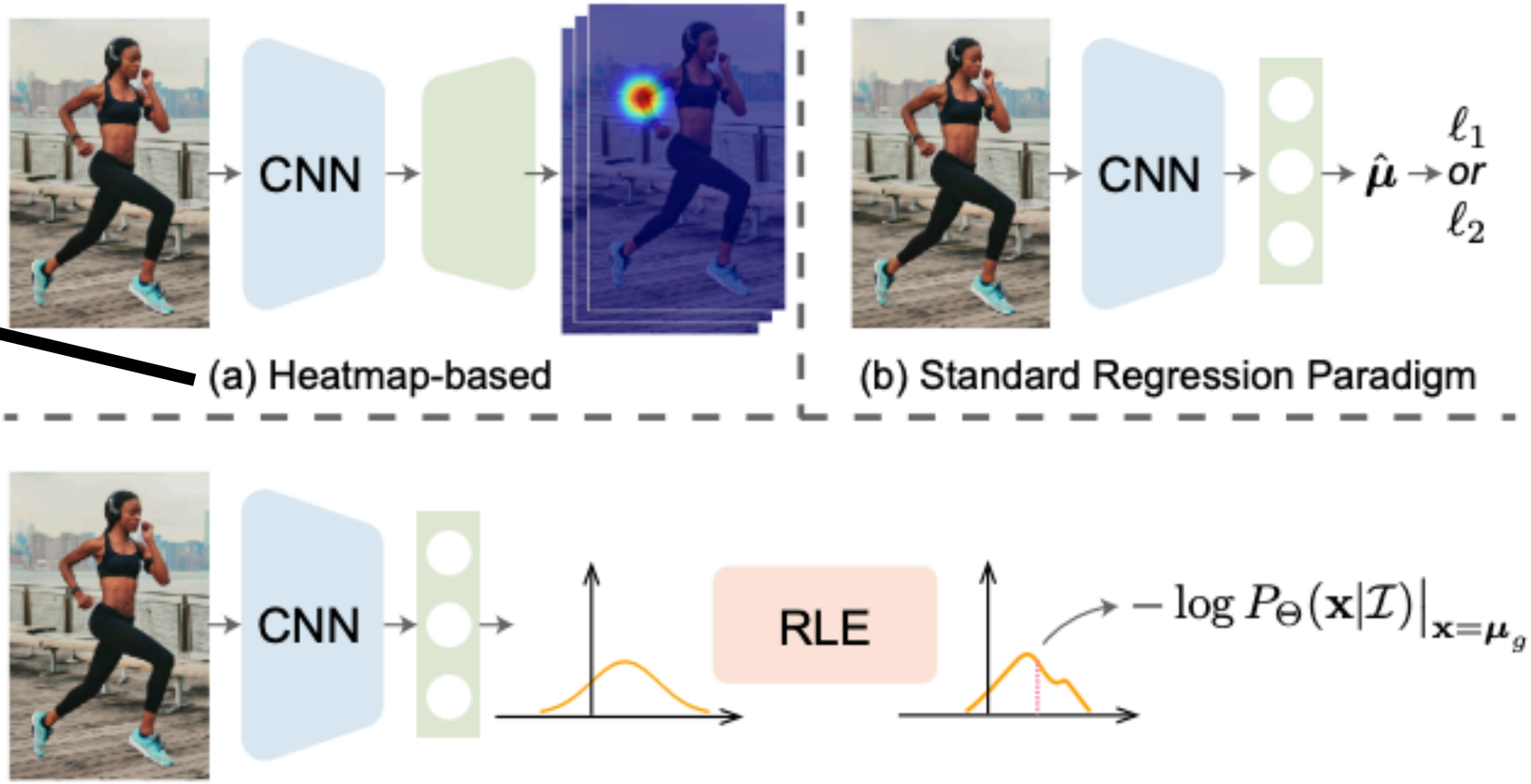
Prediction Space



Jiefeng Li et al. “Human pose regression with residual log-likelihood estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

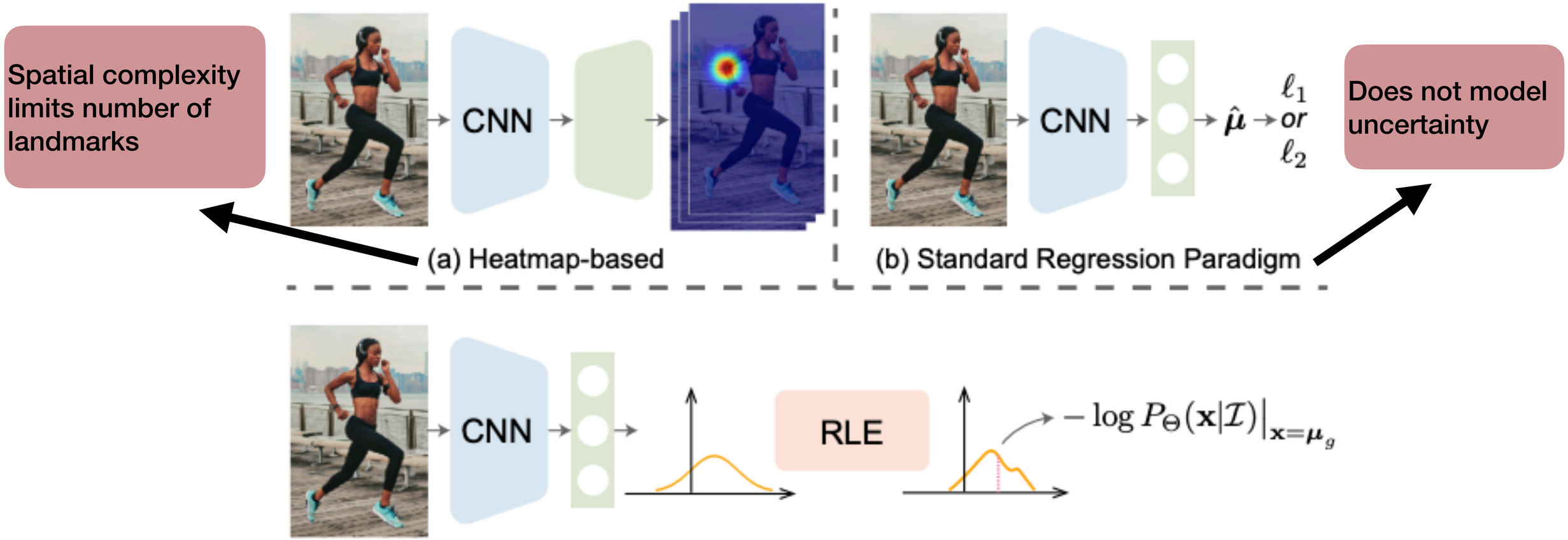
Prediction Space

Spatial complexity limits number of landmarks



Jiefeng Li et al. "Human pose regression with residual log-likelihood estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

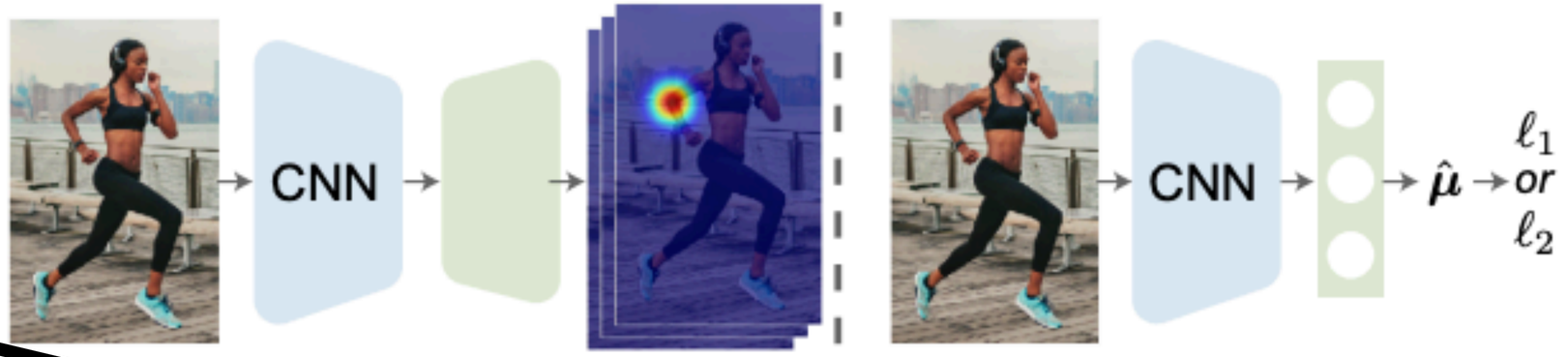
Prediction Space



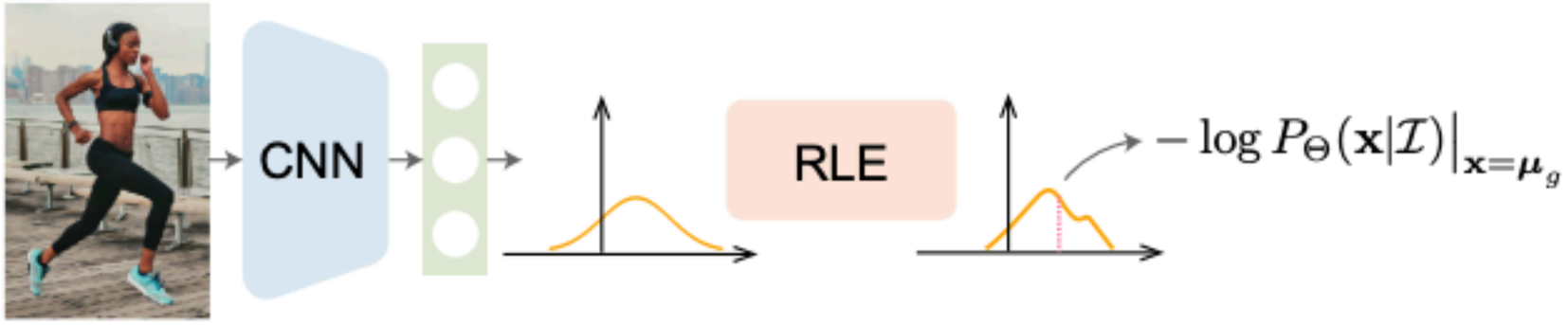
Jiefeng Li et al. “Human pose regression with residual log-likelihood estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

Prediction Space

Spatial complexity limits number of landmarks



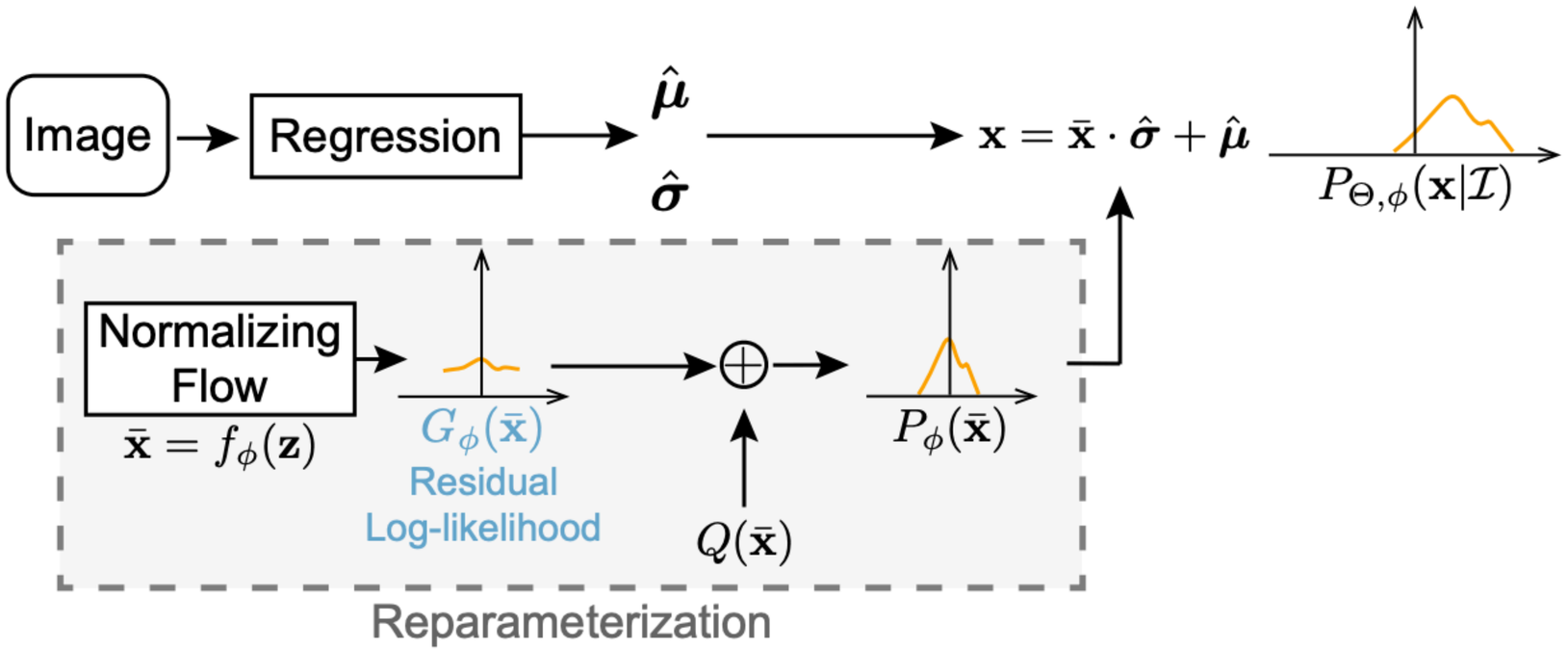
Does not model uncertainty



- Models uncertainty
- Efficient

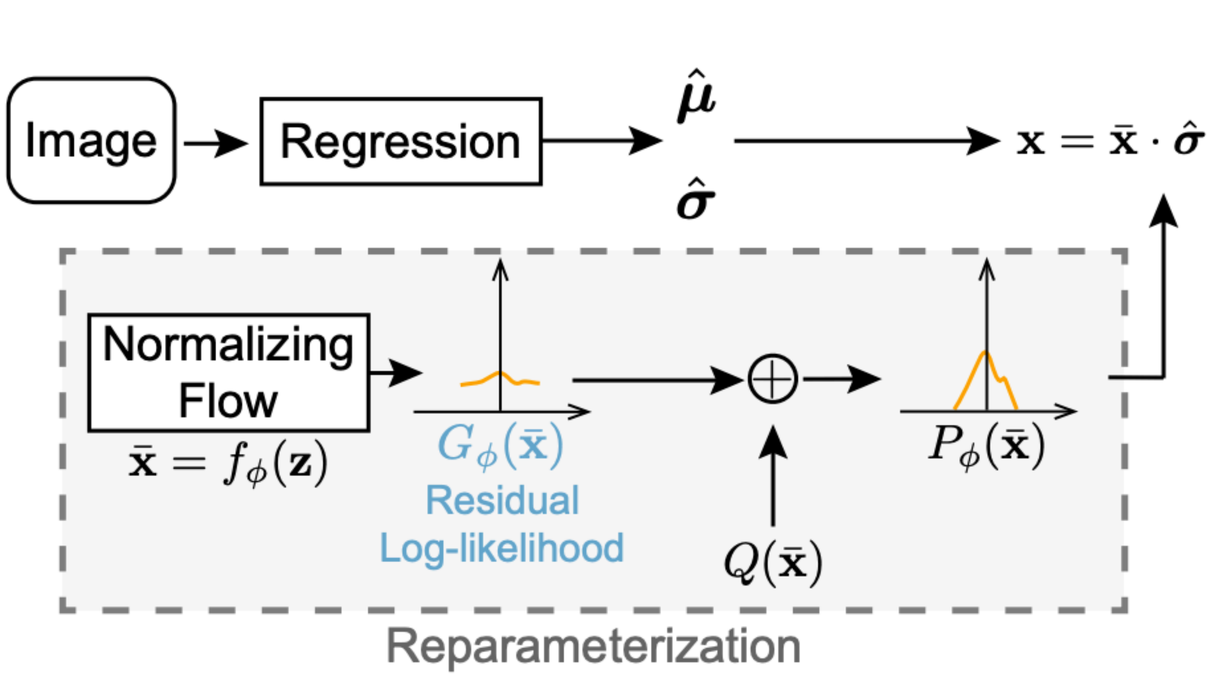
Jiefeng Li et al. "Human pose regression with residual log-likelihood estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

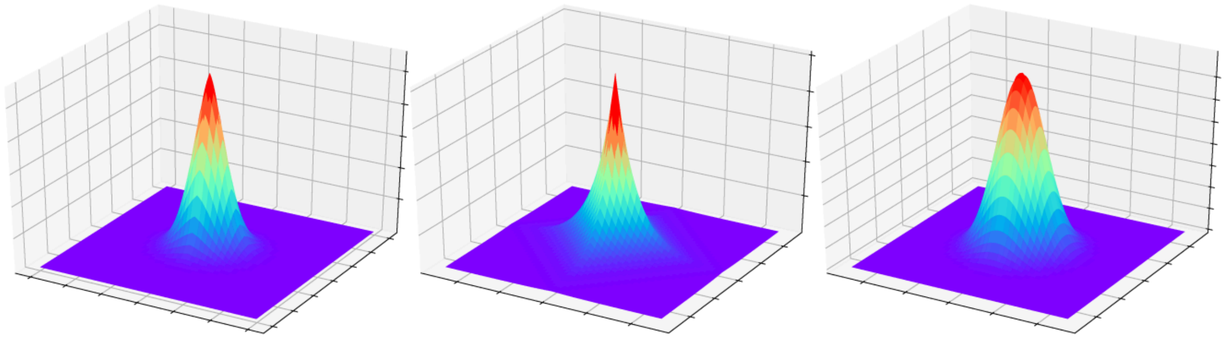
Primary Loss: Residual Log Likelihood Regression Loss (ResLL)



Jiefeng Li et al. "Human pose regression with residual log-likelihood estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

Primary Loss: Residual Log Likelihood Regression Loss (ResLL)

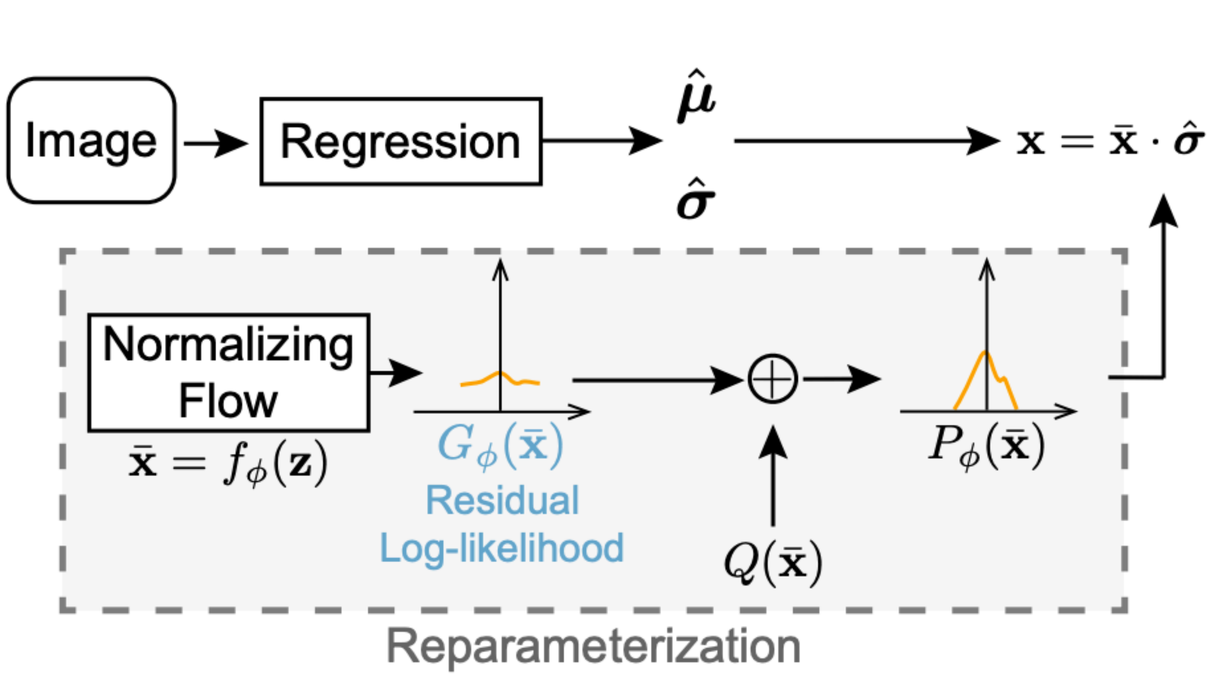


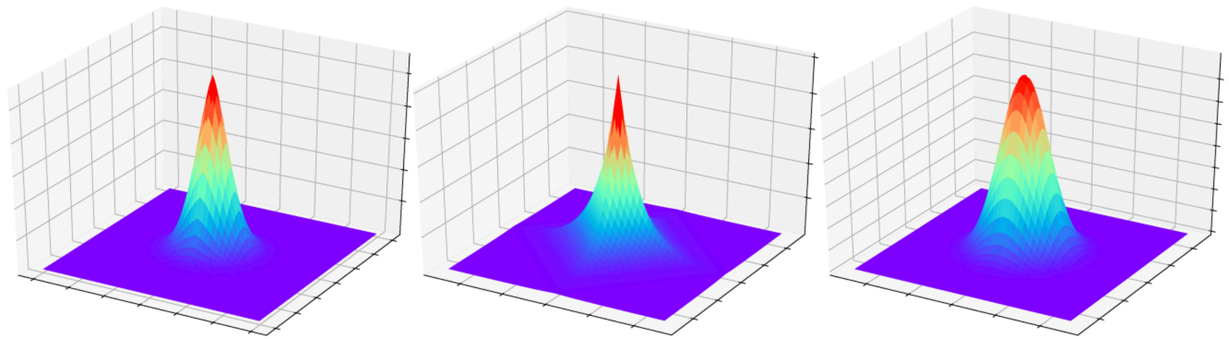


(a) The learned distribution (b) Standard Laplace distribution (c) Standard Gaussian distribution

Jiefeng Li et al. "Human pose regression with residual log-likelihood estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

Primary Loss: Residual Log Likelihood Regression Loss (ResLL)





(a) The learned distribution (b) Standard Laplace distribution (c) Standard Gaussian distribution

Landmarks

$$\hat{L} = \hat{\mu}$$

Confidences

$$\hat{C} = 1 - \hat{\sigma}$$

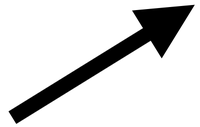
Jiefeng Li et al. "Human pose regression with residual log-likelihood estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11025–11034.

Secondary Loss: Vertex-Edge Loss (VE)

$$L_{VE} = \sum_{e \in E} \left(\hat{e} - e_{gt} \right)^2$$

Secondary Loss: Vertex-Edge Loss (VE)

$$L_{VE} = \sum_{e \in E} \left(\hat{e} - e_{gt} \right)^2$$



Enforce face geometry

Model Fitting

Reprojection Loss

$$L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \Psi); \Pi) = \frac{1}{2} \, || \, \hat{C} \; r(\hat{L}, V_L(\beta, \theta, \Psi); \Pi) \, ||^2$$

Model Fitting

$$\hat{L} - \Pi(V_L(\beta, \theta, \Psi); K)$$

Reprojection Loss

$$L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \Psi); \Pi) = \frac{1}{2} \, || \hat{C} \; r(\hat{L}, V_L(\beta, \theta, \Psi); \Pi) \, ||^2$$

Model Fitting

$$\hat{L} - \Pi(V_L(\beta, \theta, \Psi); K)$$

Reprojection Loss

$$L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \Psi); \Pi) = \frac{1}{2} \, || \hat{C} \; r(\hat{L}, V_L(\beta, \theta, \Psi); \Pi) \, ||^2$$

$$\begin{pmatrix} \sqrt{H^2 + W^2} & 0 & W/2 \\ 0 & \sqrt{H^2 + W^2} & H/2 \\ 0 & 0 & 1 \end{pmatrix}$$

Imry Kissos et al. “Beyond weak perspective for monocular 3D human pose estimation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 541–554.

Model Fitting

$$\hat{L} - \Pi(V_L(\beta, \theta, \Psi); K)$$

Reprojection Loss

$$L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \Psi); \Pi) = \frac{1}{2} \, || \hat{C} \; r(\hat{L}, V_L(\beta, \theta, \Psi); \Pi) \, ||^2$$

$$\begin{pmatrix} \sqrt{H^2 + W^2} & 0 & W/2 \\ 0 & \sqrt{H^2 + W^2} & H/2 \\ 0 & 0 & 1 \end{pmatrix}$$

OPTIMIZATION

(1) Rigid Fitting

$$\arg\min_{\theta_r} L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \psi); \Pi)$$

(2) Non Rigid Fitting

$$\arg\min_{\beta, \theta, \psi} L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \psi); \Pi)$$

Imry Kissos et al. “Beyond weak perspective for monocular 3D human pose estimation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 541–554.

Model Fitting

$$\hat{L} - \Pi(V_L(\beta, \theta, \Psi); K)$$

Reprojection Loss

$$L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \Psi); \Pi) = \frac{1}{2} \, || \hat{C} \; r(\hat{L}, V_L(\beta, \theta, \Psi); \Pi) \, ||^2$$

$$\begin{pmatrix} \sqrt{H^2 + W^2} & 0 & W/2 \\ 0 & \sqrt{H^2 + W^2} & H/2 \\ 0 & 0 & 1 \end{pmatrix}$$

OPTIMIZATION

(1) Rigid Fitting

$$\arg\min_{\theta_r} L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \psi); \Pi)$$

(2) Non Rigid Fitting

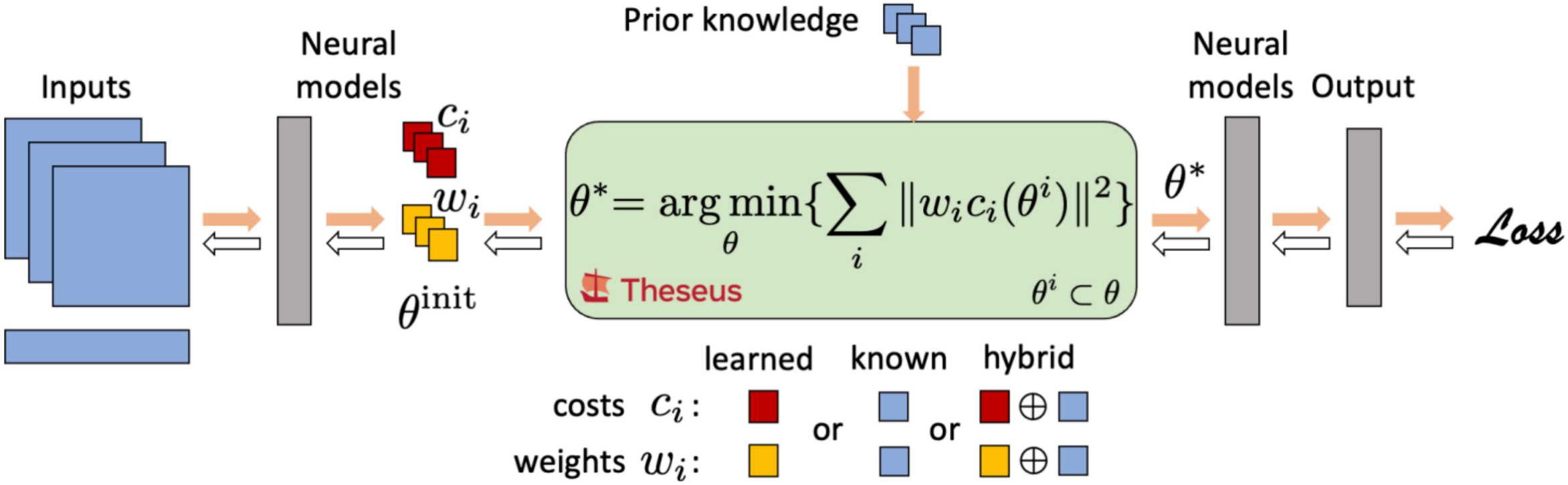
$$\arg\min_{\beta, \theta, \psi} L_{RP}(\hat{L}, \hat{C}, V_L(\beta, \theta, \psi); \Pi)$$

Non Linerar Least Squares

- Gradient Descent
- Gauss-Newton
- Levenberg-Marquardt

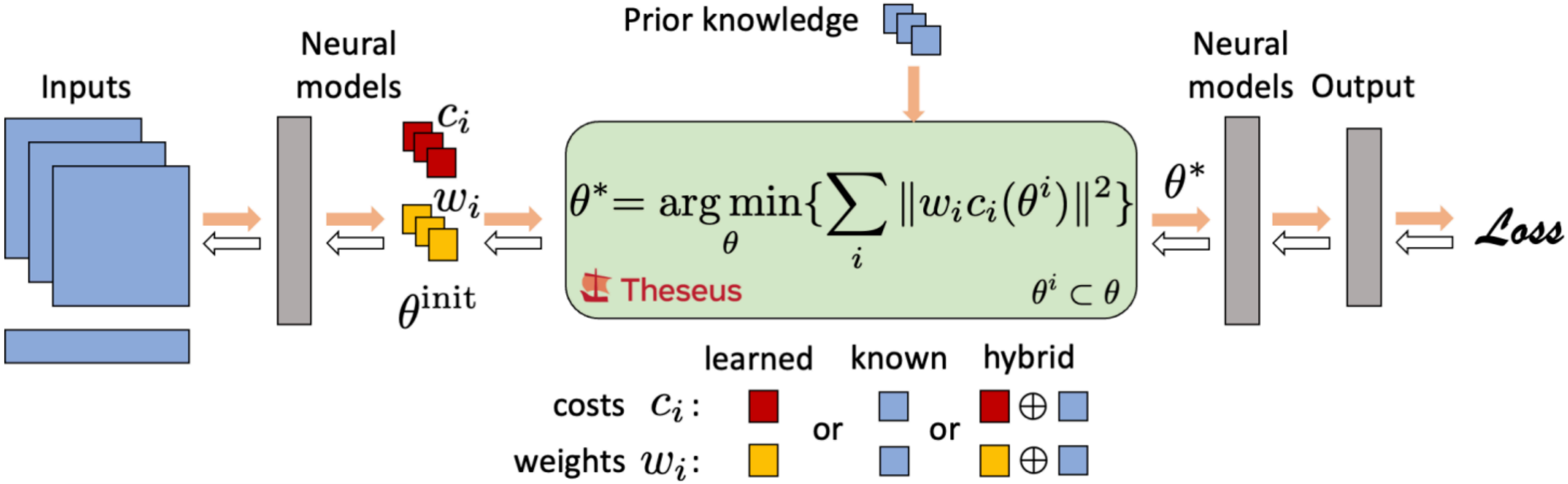
Imry Kissos et al. “Beyond weak perspective for monocular 3D human pose estimation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 541–554.

Differentiable Non Linear Optimization: Theseus Library



Luis Pineda et al. “Theseus: A Library for Differentiable Nonlinear Optimization”. In: *Advances in Neural Information Processing Systems* (2022).

Differentiable Non Linear Optimization: Theseus Library



Note: we only use it for the optimisation

Luis Pineda et al. “Theseus: A Library for Differentiable Nonlinear Optimization”. In: *Advances in Neural Information Processing Systems* (2022).

Quantitative Results: Loss

$$\frac{1}{NL}\sum_{i=1}^{N}\sum_{k=1}^{L}||\hat{L}_{ik} - L_{ik}||_2$$

Loss	VE Loss	MLME (px) ↓
NGLL	✗	2.17
NGLL	✓	2.18
ResLL	✗	2.20
ResLL	✓	2.18
ResLL Exp	✗	**2.11**
ResLL Exp	✓	2.13

Quantitative Results: Loss

$$2 \log \hat{\sigma} + \frac{(\mu_{gt} - \hat{\mu})^2}{2\hat{\sigma}^2}$$

Used by Wood et al.

$$\frac{1}{NL} \sum_{i=1}^{N} \sum_{k=1}^{L} || \hat{L}_{ik} - L_{ik} ||_2$$

Loss	VE Loss	MLME (px) ↓
NGLL	✗	2.17
NGLL	✓	2.18
ResLL	✗	2.20
ResLL	✓	2.18
ResLL Exp	✗	**2.11**
ResLL Exp	✓	2.13

Quantitative Results: Loss

$$2\log\hat{\sigma} + \frac{(\mu_{gt} - \hat{\mu})^2}{2\hat{\sigma}^2}$$

Used by Wood et al.

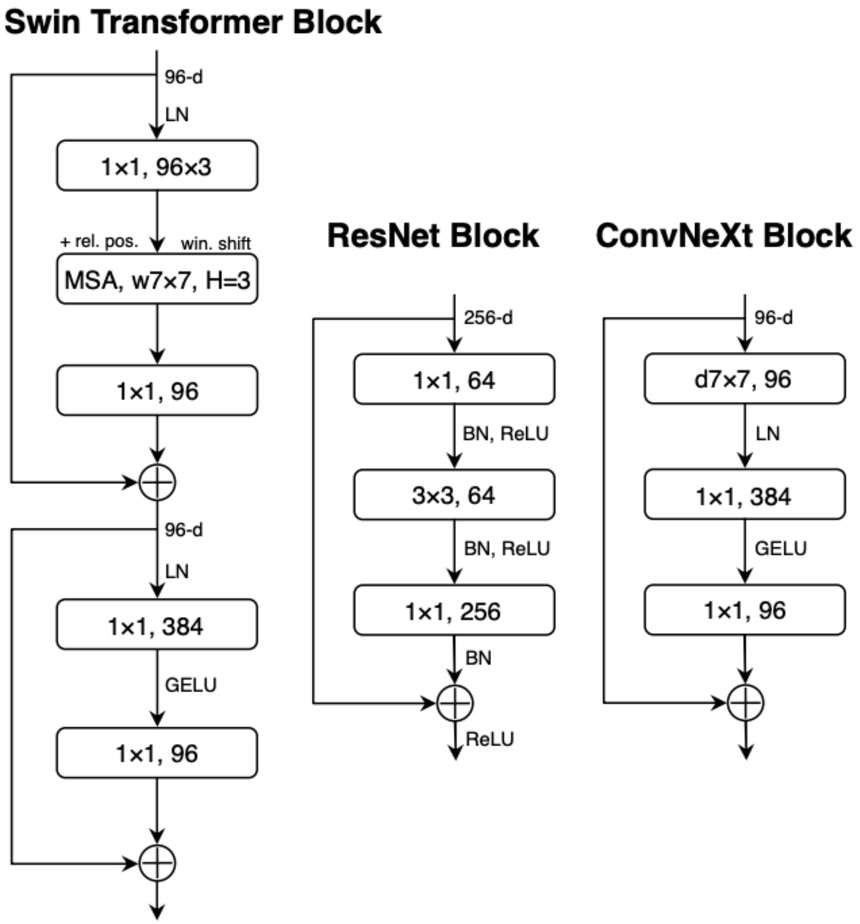
$$\frac{1}{NL}\sum_{i=1}^{N}\sum_{k=1}^{L}||\hat{L}_{ik} - L_{ik}||_2$$

Loss	VE Loss	MLME (px) ↓
NGLL	✗	2.17
NGLL	✓	2.18
ResLL	✗	2.20
ResLL	✓	2.18
ResLL Exp	✗	**2.11**
ResLL Exp	✓	2.13

ResLL
Best accuracy

Quantitative Results: Predictor Backbone

Backbone	VE Loss	MLME (px) ↓
ResNet-50	✗	2.11
ResNet-50	✓	2.13
ConvNeXt Tiny	✗	2.13
ConvNeXt Tiny	✓	**2.07**
Swin Transformer Tiny	✗	2.38
Swin Transformer Tiny	✓	2.35
MobileNet V3 Large	✗	2.55
MobileNet V3 Large	✓	2.57



Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.

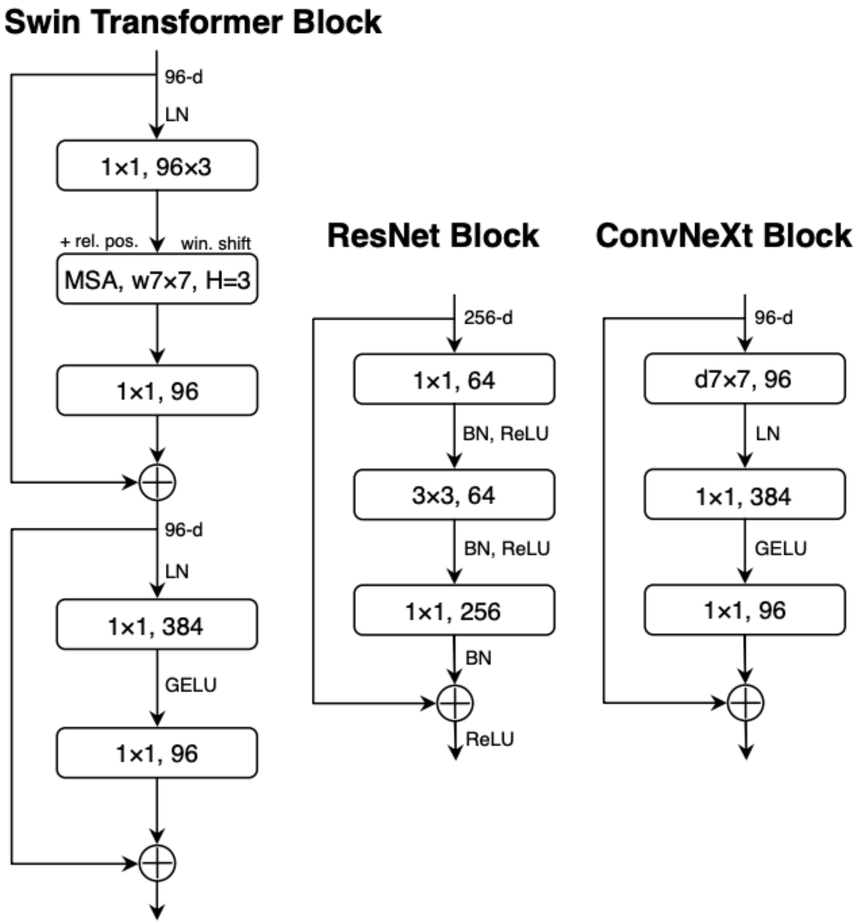
Andrew Howard et al. “Searching for mobilenetv3”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

Quantitative Results: Predictor Backbone

Backbone	VE Loss	MLME (px) ↓
ResNet-50	✗	2.11
ResNet-50	✓	2.13
ConvNeXt Tiny	✗	2.13
ConvNeXt Tiny	✓	**2.07**
Swin Transformer Tiny	✗	2.38
Swin Transformer Tiny	✓	2.35
MobileNet V3 Large	✗	2.55
MobileNet V3 Large	✓	2.57

ConvNeXt + Ve

Best accuracy



Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.

Andrew Howard et al. “Searching for mobilenetv3”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

Quantitative Results: Predictor Decoder

$$\frac{1}{NV_S}\sum_{i=1}^{N}\sum_{k=1}^{V_S}||\hat{V}_{S,ik} - V_{S,ik}||_2$$

Backbone	GNN Decoder	VE Loss	MLME (px) ↓	MVE (mm) ↓ Face	Head	MRLE (px)↓
ResNet-50	✗	✗	2.11	07.9	10.7	**0.48**
ResNet-50	✗	✓	2.13	08.2	11.2	0.49
ResNet-50	✓	✗	2.26	08.0	11.2	0.76
ResNet-50	✓	✓	**2.01**	08.1	10.9	0.51
ConvNeXt Tiny	✗	✗	2.13	07.9	10.8	0.53
ConvNeXt Tiny	✗	✓	2.07	08.0	11.0	0.49
ConvNeXt Tiny	✓	✗	2.10	**07.8**	**10.6**	0.51
ConvNeXt Tiny	✓	✓	2.18	08.1	11.1	0.52

Quantitative Results: Predictor Decoder

$$\frac{1}{NV_S}\sum_{i=1}^{N}\sum_{k=1}^{V_S}||\hat{V}_{S,ik} - V_{S,ik}||_2$$

Backbone	GNN Decoder	VE Loss	MLME (px) ↓	MVE (mm) ↓ Face	Head	MRLE (px)↓
ResNet-50	✗	✗	2.11	07.9	10.7	**0.48**
ResNet-50	✗	✓	2.13	08.2	11.2	0.49
ResNet-50	✓	✗	2.26	08.0	11.2	0.76
ResNet-50	✓	✓	**2.01**	08.1	10.9	0.51
ConvNeXt Tiny	✗	✗	2.13	07.9	10.8	0.53
ConvNeXt Tiny	✗	✓	2.07	08.0	11.0	0.49
ConvNeXt Tiny	✓	✗	2.10	**07.8**	**10.6**	0.51
ConvNeXt Tiny	✓	✓	2.18	08.1	11.1	0.52

Graph Decoder

- Best accuracy 2D
- Best accuracy 3D

Quantitative Results: Predictor Decoder

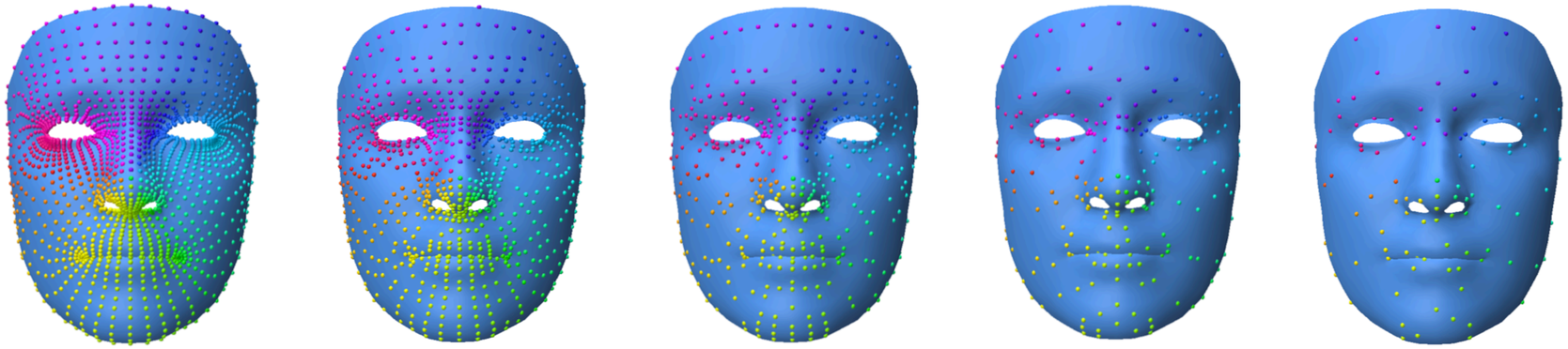
$$\frac{1}{NV_S}\sum_{i=1}^{N}\sum_{k=1}^{V_S}||\hat{V}_{S,ik} - V_{S,ik}||_2$$

Backbone	GNN Decoder	VE Loss	MLME (px) ↓	MVE (mm) ↓ Face	Head	MRLE (px)↓
ResNet-50	✗	✗	2.11	07.9	10.7	**0.48**
ResNet-50	✗	✓	2.13	08.2	11.2	0.49
ResNet-50	✓	✗	2.26	08.0	11.2	0.76
ResNet-50	✓	✓	**2.01**	08.1	10.9	0.51
ConvNeXt Tiny	✗	✗	2.13	07.9	10.8	0.53
ConvNeXt Tiny	✗	✓	2.07	08.0	11.0	0.49
ConvNeXt Tiny	✓	✗	2.10	**07.8**	**10.6**	0.51
ConvNeXt Tiny	✓	✓	2.18	08.1	11.1	0.52

Graph Decoder

- Best accuracy 2D
- Best accuracy 3D

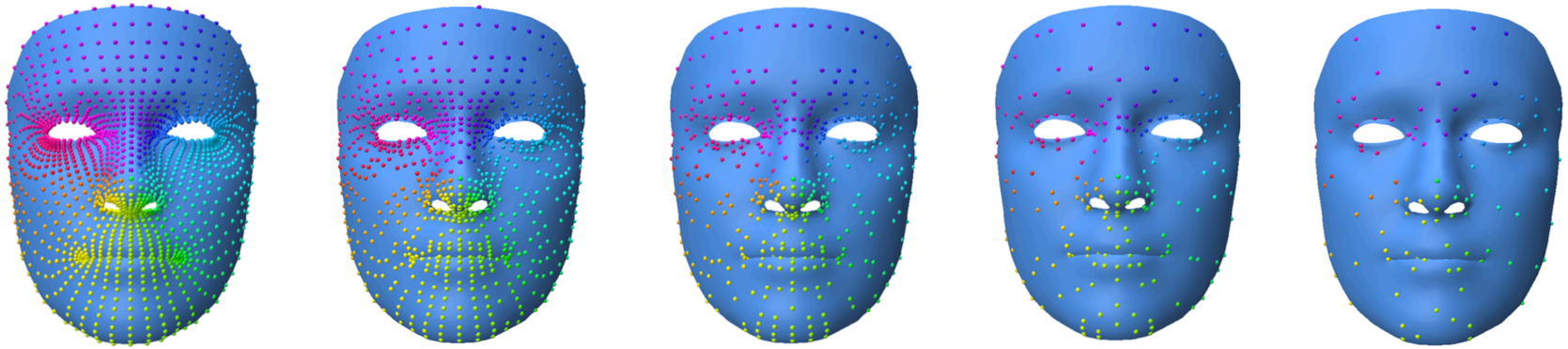
Quantitative Results: Different Landmark Sets for Fitting



1609 805 403 202 101

Subset	MVE (mm) ↓ Face	Head	Time (s/I)
1609	07.8	10.6	0.9
805	**07.5**	**10.4**	1.0
403	07.5	10.6	1.1
202	07.6	10.8	1.1
101	07.8	10.8	1.2

Quantitative Results: Different Landmark Sets for Fitting

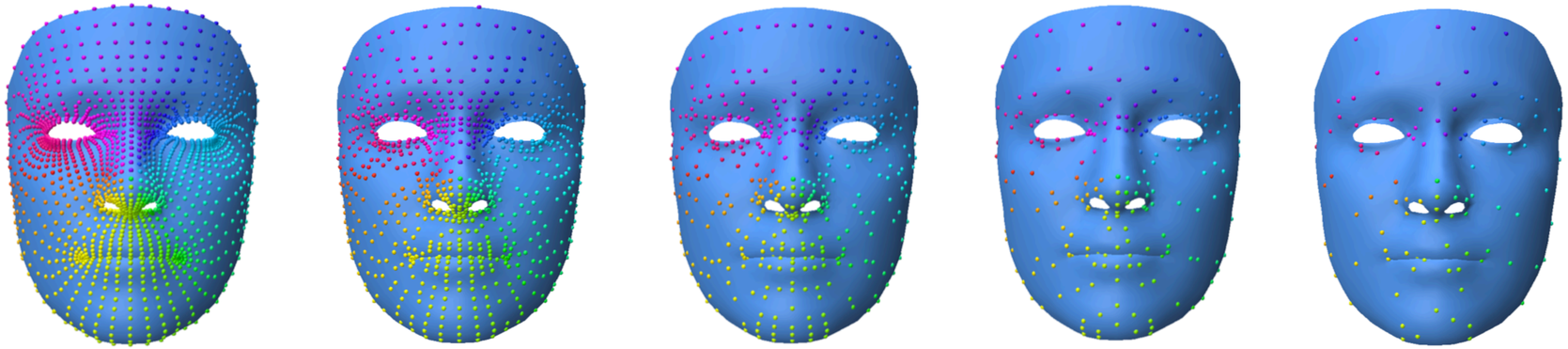


1609 805 403 202 101

805 landmarks
Best accuracy

Subset	MVE (mm) ↓ Face	Head	Time (s/I)
1609	07.8	10.6	0.9
805	**07.5**	**10.4**	1.0
403	07.5	10.6	1.1
202	07.6	10.8	1.1
101	07.8	10.8	1.2

Quantitative Results: Different Landmark Sets for Fitting



1609 805 403 202 101

805 landmarks

Best accuracy

Subset	MVE (mm) ↓ Face	Head	Time (s/I)
1609	07.8	10.6	0.9
805	**07.5**	**10.4**	1.0
403	07.5	10.6	1.1
202	07.6	10.8	1.1
101	07.8	10.8	1.2

Dropping Landmarks

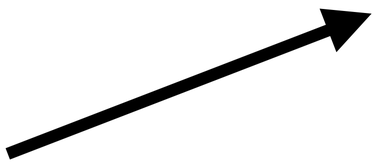
- Accuracy slowly drops
- Times increase

Quantitative Results: DAD-3DNet, Deca comparison over DAD-3DHead images

Method	MVE (mm) ↓ Face	Head	Time (s/I)
DAD-3DNet	11.8	18.2	0.6
DECA	11.3	16.0	0.1
ours	**07.5**	**10.4**	1.0

Quantitative Results: DAD-3DNet, Deca comparison over DAD-3DHead images

Method	MVE (mm) ↓ Face	Head	Time (s/I)
DAD-3DNet	11.8	18.2	0.6
DECA	11.3	16.0	0.1
ours	**07.5**	**10.4**	1.0

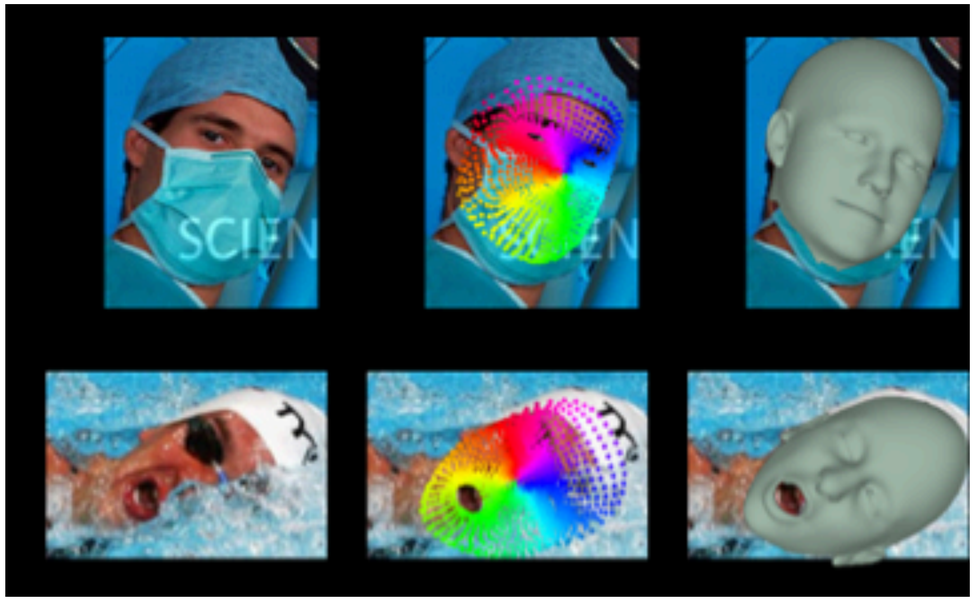


Best Accuracy

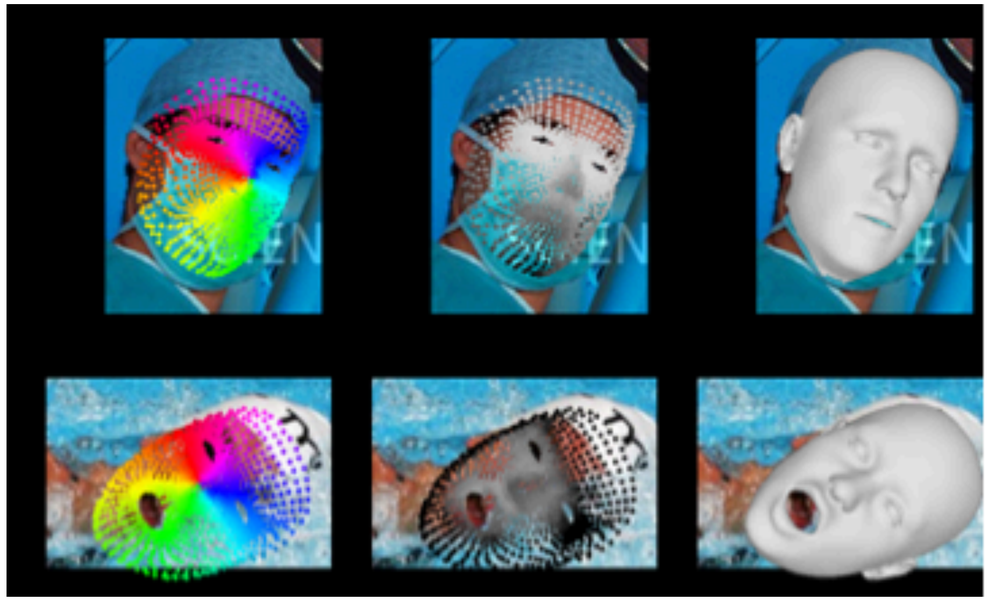
Slowest

Qualitative Results from Test Set

Pseudo-Ground Truth



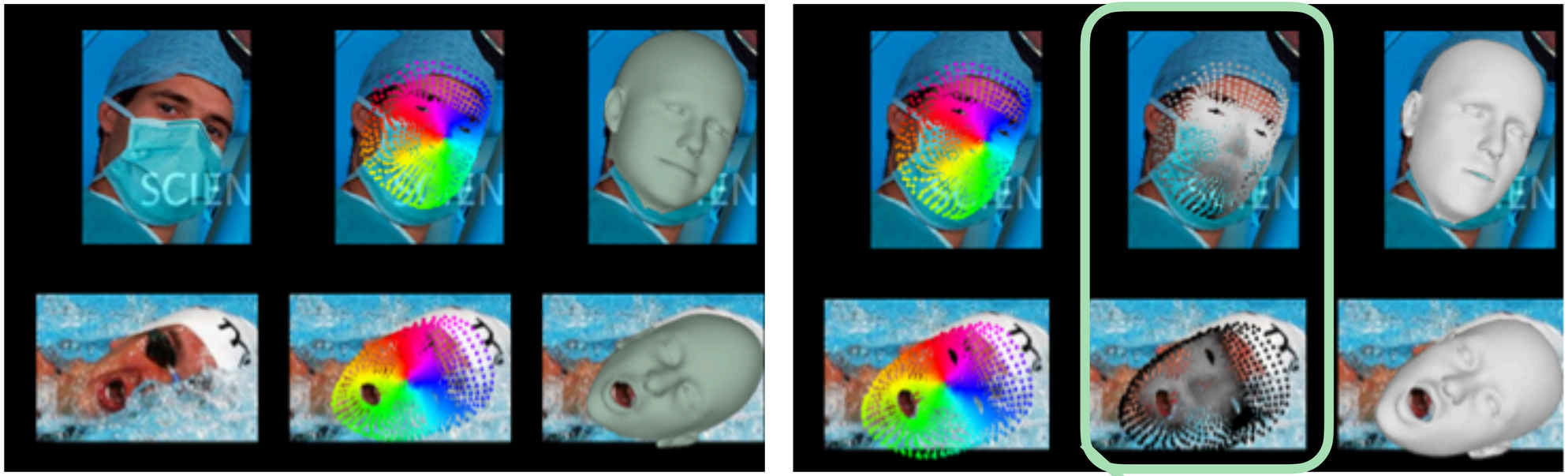
Prediction



Qualitative Results from Test Set

Pseudo-Ground Truth

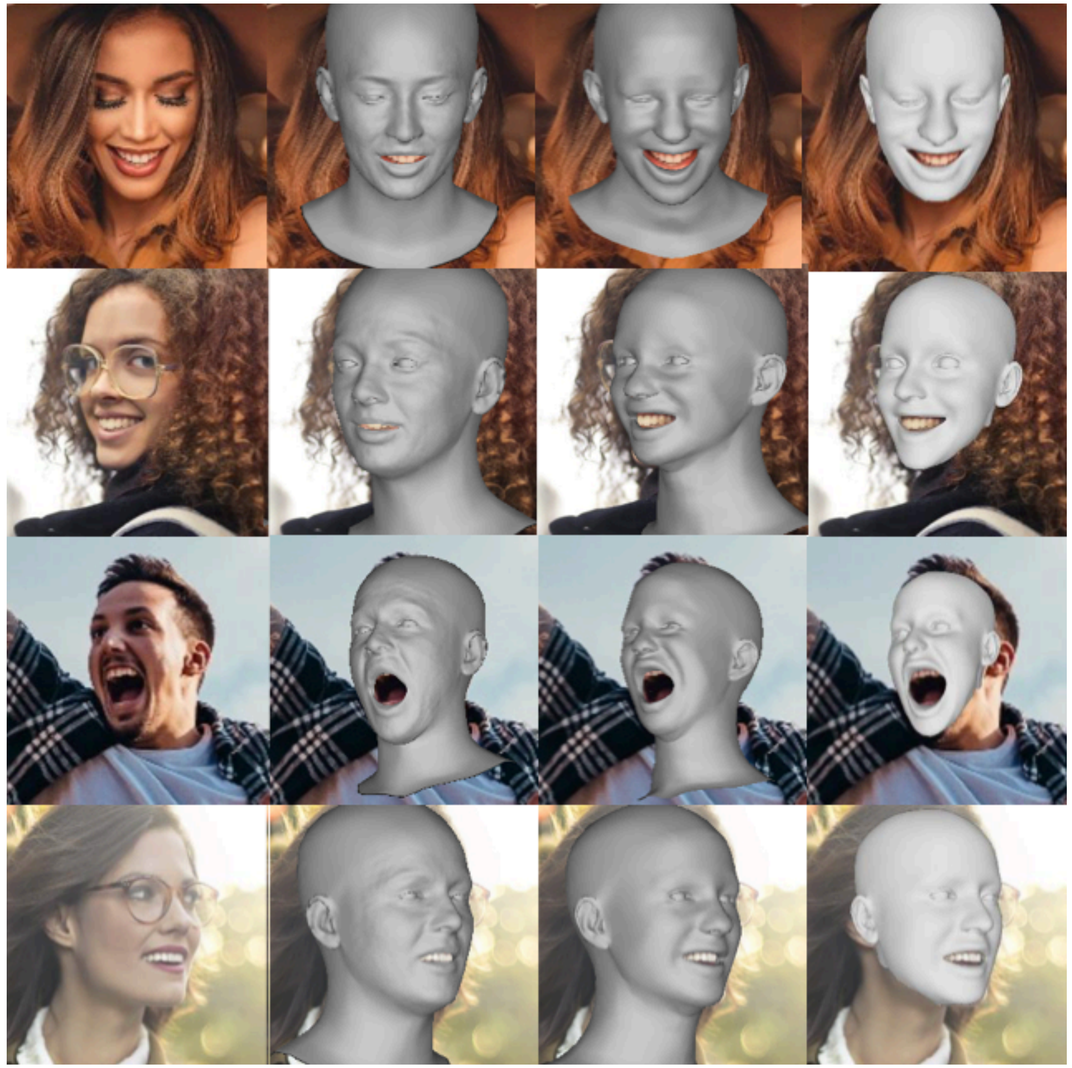
Prediction



Meaningful Confidences

Qualitative Comparison

DECA DAD-3DNet OURS



DECA DAD-3DNet OURS

Qualitative Comparison

DECA DAD-3DNet OURS

DECA DAD-3DNet OURS

Conclusion

Current Limitations

- Poor initialization for the fitting
- Predictor is over-confident
- Confidence values are not spread out in $[0,1]$
- Head alignment issue

Conclusion

Current Limitations

- Poor initialization for the fitting
- Predictor is over-confident
- Confidence values are not spread out in $[0,1]$

- Head alignment issue

Future Work

- End to end predictor with differentiable optimisation

- Further analysis on best set of landmarks

- Use it for hand reconstruction

Conclusion

Current Limitations

- Poor initialization for the fitting
- Predictor is over-confident
- Confidence values are not spread out in $[0,1]$
- Head alignment issue

Future Work

- End to end predictor with differentiable optimisation
- Further analysis on best set of landmarks
- Use it for hand reconstruction

Thank You!