Unsupervised Discovery of Object Landmarks as Structural Representations

Yuting Zhang[1], Yijie Guo[1], Yixin Jin[1], Yijun Luo[1], Zhiyuan He[1], Honglak Lee[1,2]

[1] University of Michigan, Ann Arbor

[2] Google Brain

Structural representations of images

- Computer vision seeks to understand visual structures.
  - Poses, contours, 3D shapes, …
  - Physically conceptualized, perceptible by humans
- Deep neural networks can learn latent representations.
  - Desired properties: distributed, sparse, transferable, …
  - Not as conceptualized and interpretable as explicit structures

Structural representations of images

- Computer vision seeks to understand visual structures.
  - Poses, contours, 3D shapes, …
  - Physically conceptualized, perceptible by humans
- Deep neural networks can learn latent representations.
  - Desired properties: distributed, sparse, transferable, …
  - Not as conceptualized and interpretable as explicit structures
- Typically, extra supervision is needed to bridge the gap between latent representations and explicit structures
  - costly to obtain and often unavailable

Can we train a deep neural network to get image **representations of explicit structures without supervision**?

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

The explicit structure

Can we train a deep neural network to get image **representations of explicit structures without supervision**?

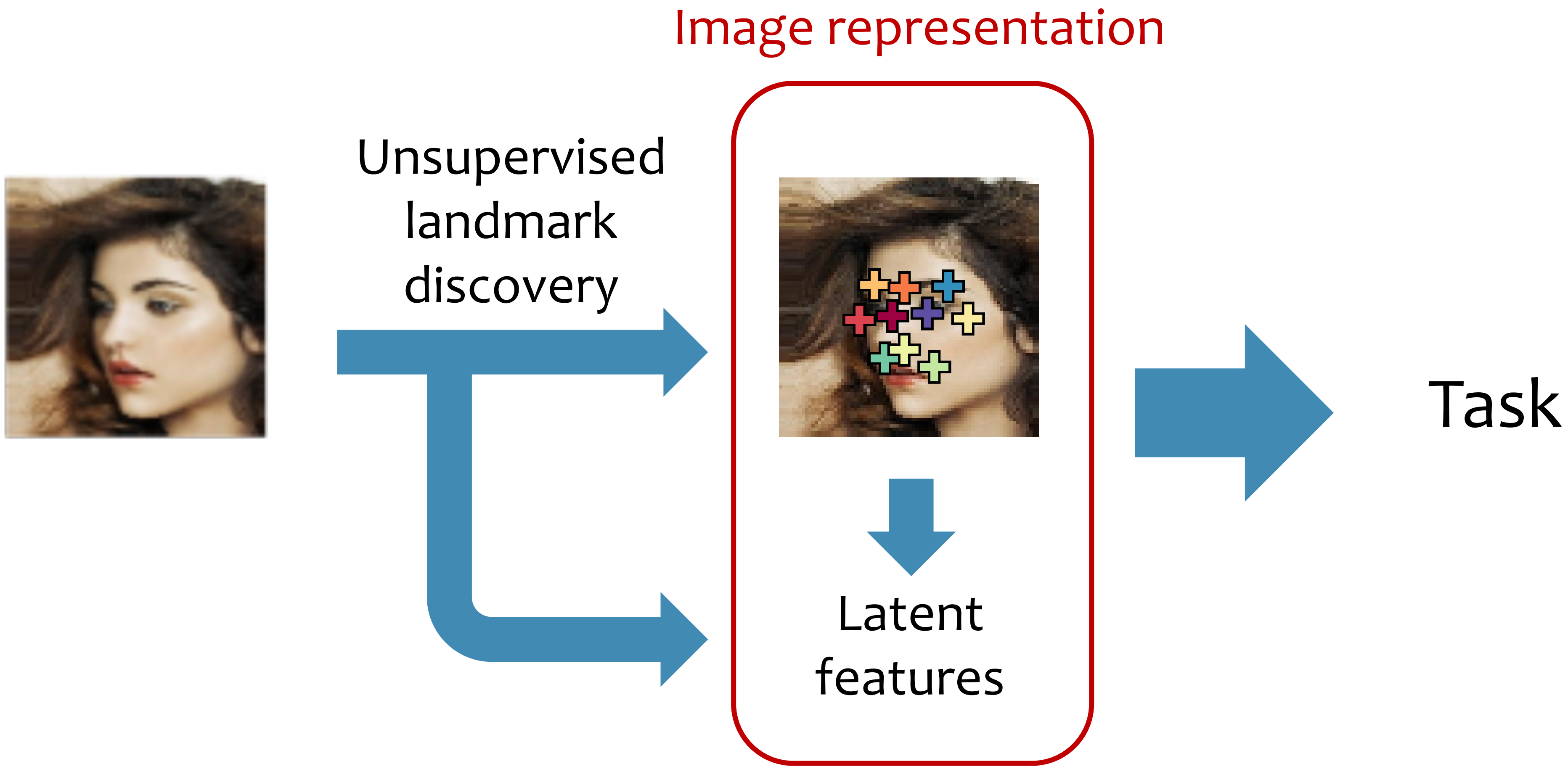
- We consider a specific type of explicit structures:

Object landmarks

- Compact representation of object shapes
- Generally applicable to many object categories

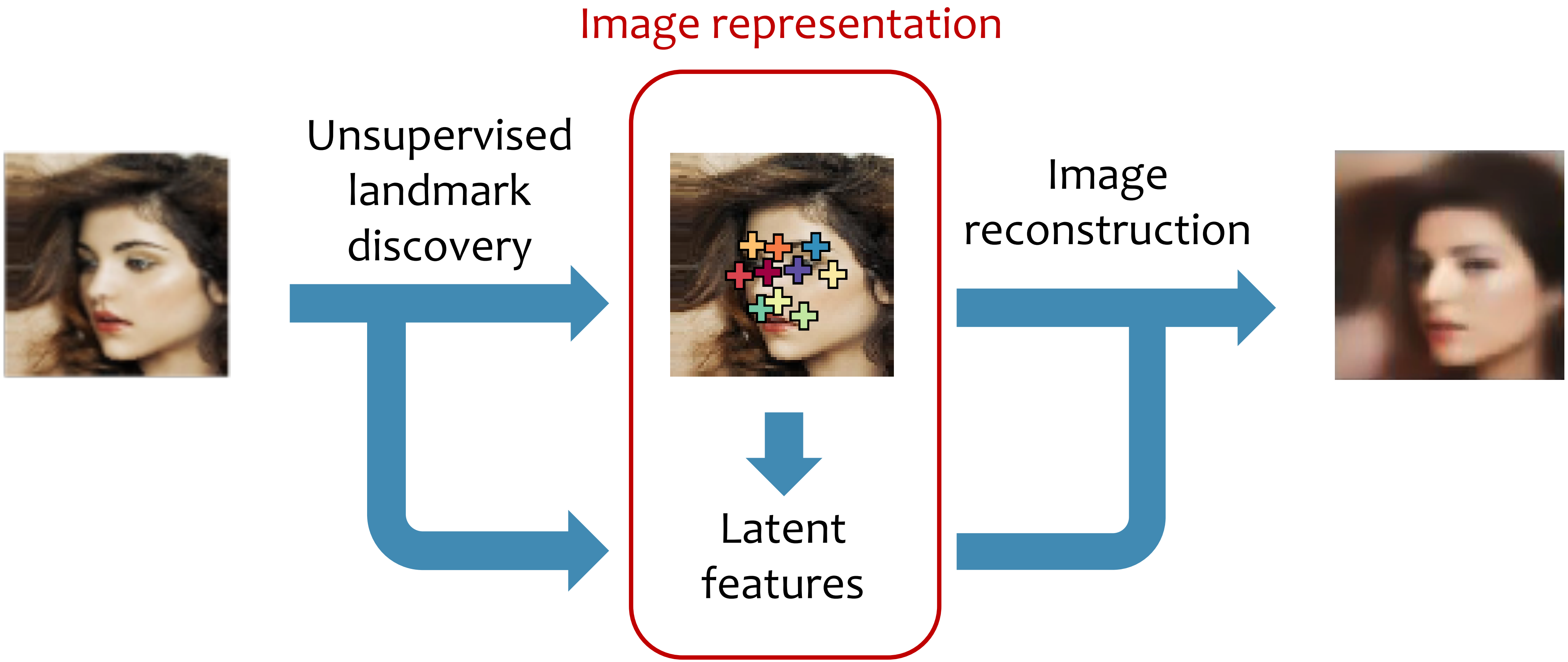
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Our framework



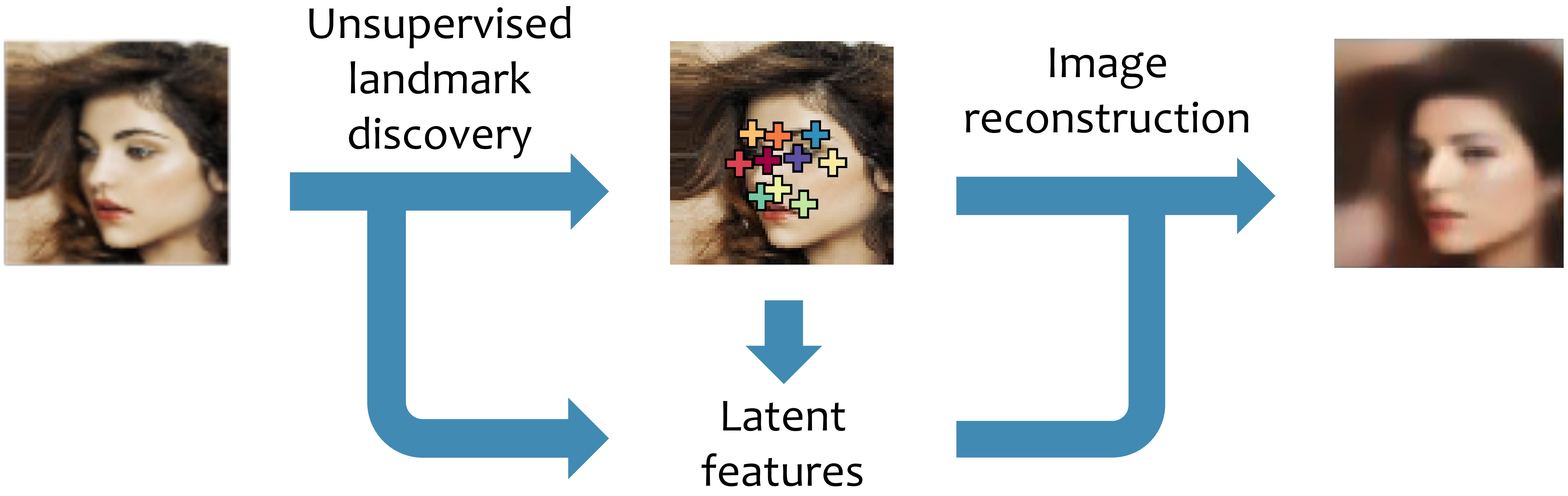
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Our framework



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

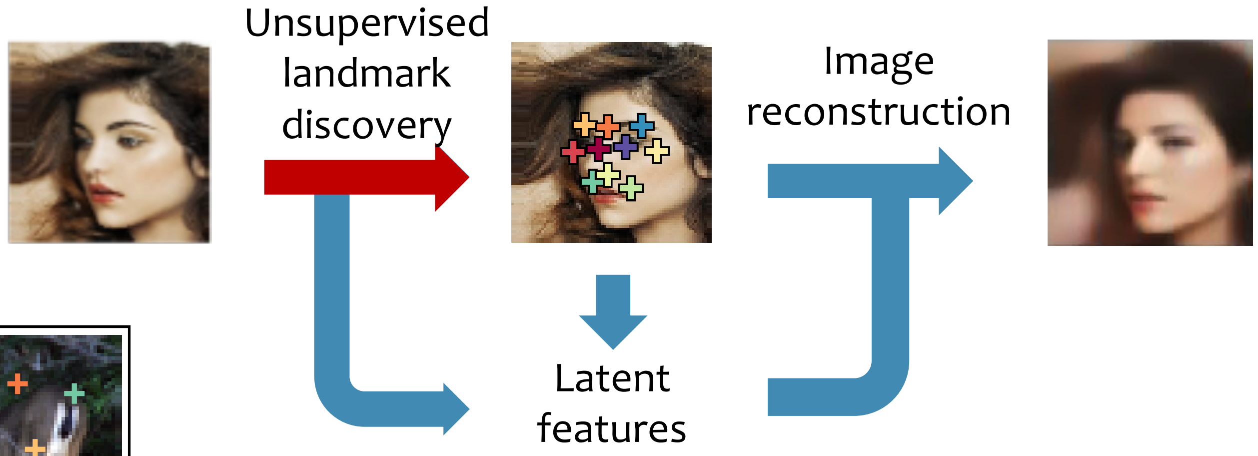
Our framework



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

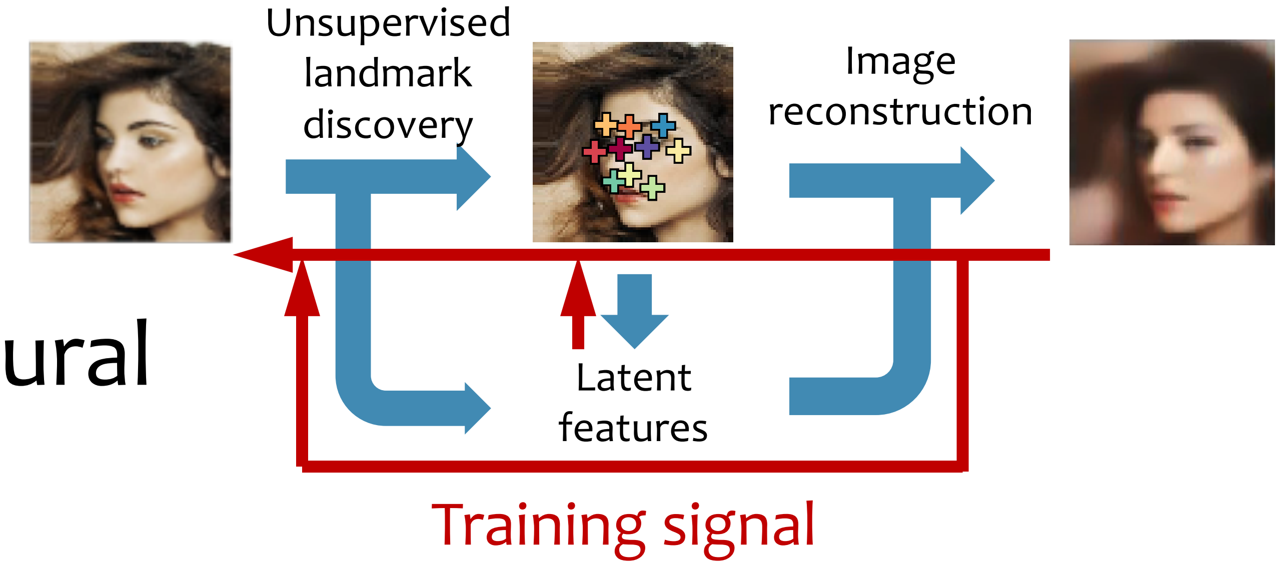
Technical outline

- Unsupervised object landmark discovery



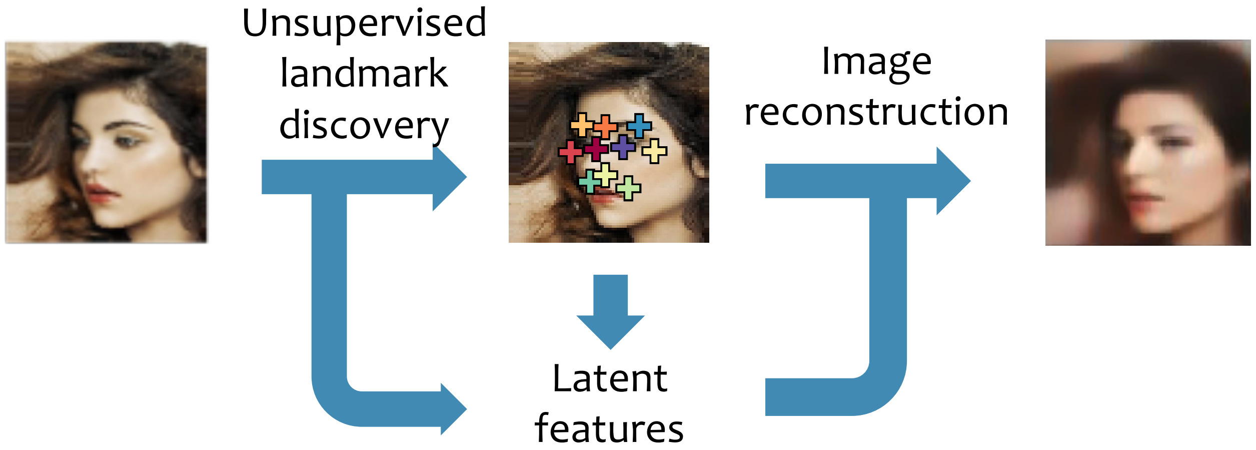
Technical outline

- Unsupervised object landmark discovery
- A fully differentiable neural network architecture



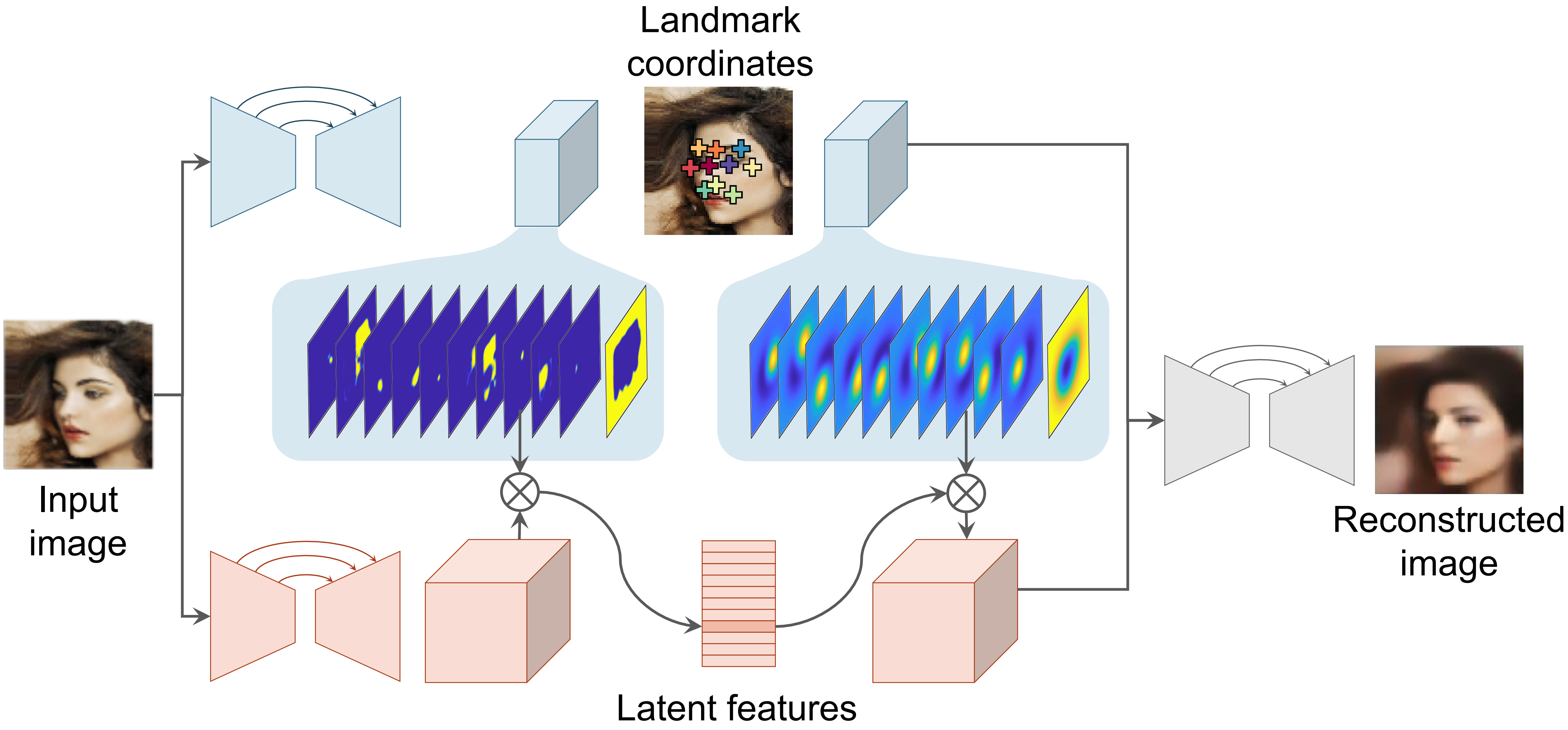
- The image reconstruction can encourage the learning of informative landmarks and features.

Technical outline



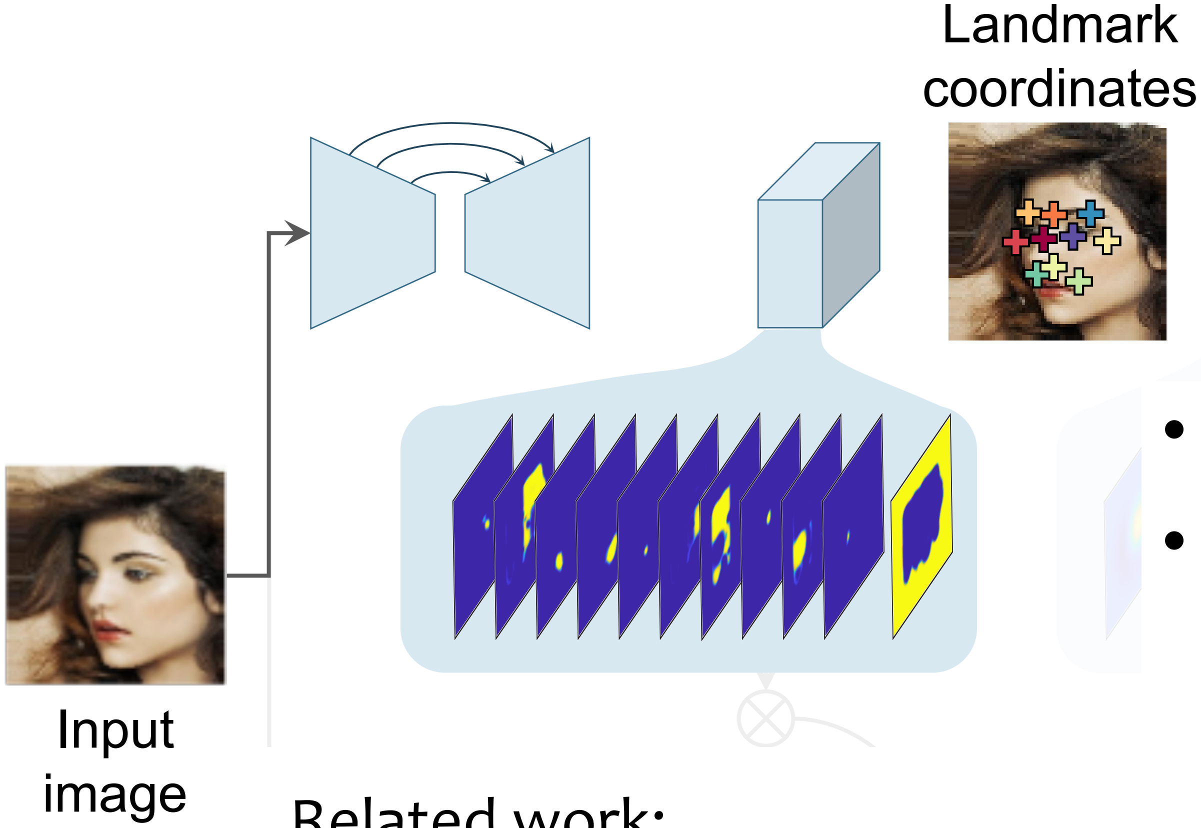
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture



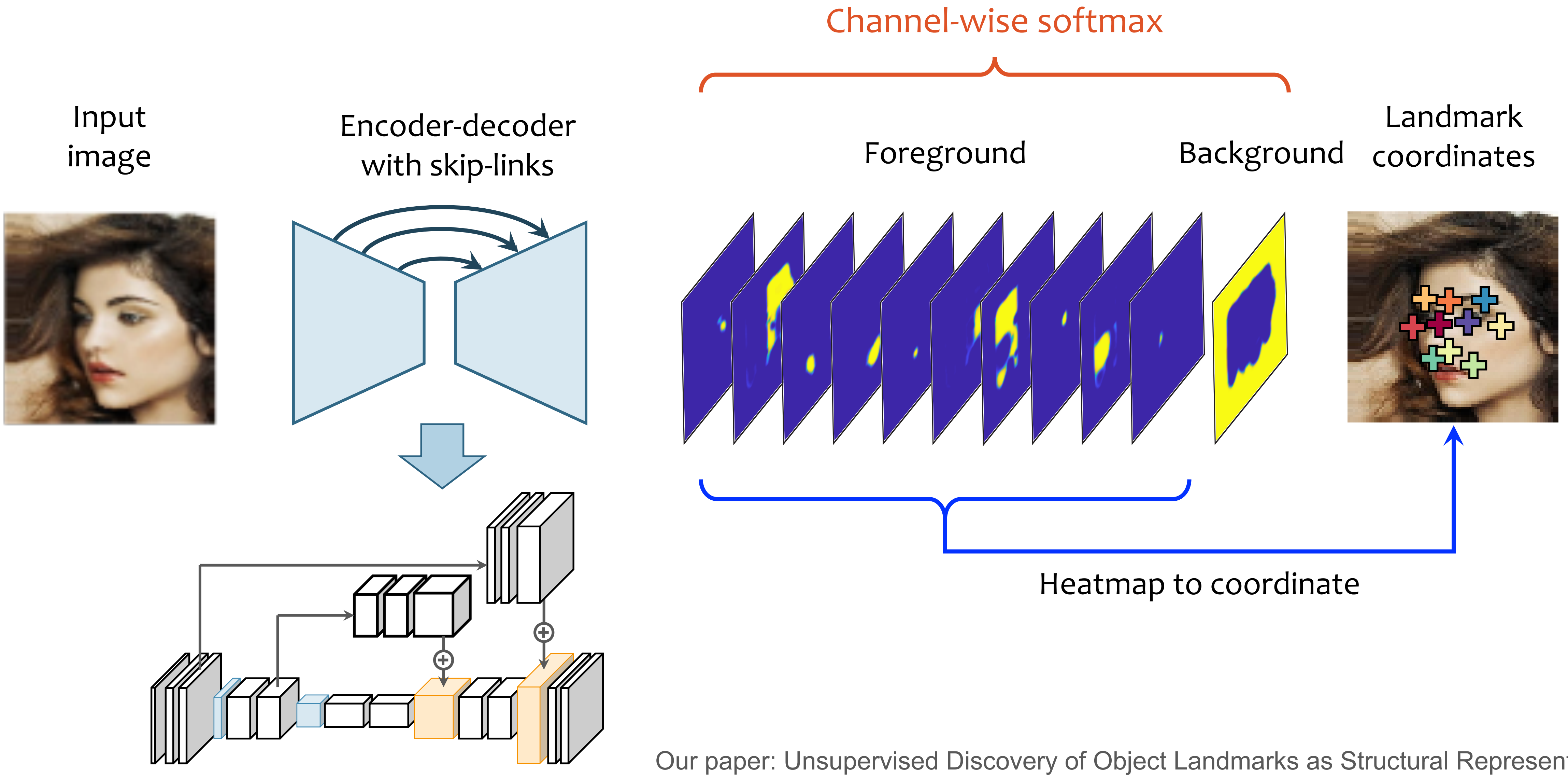
Unsupervised landmark discovery

- A differentiable formulation
- Unsupervised constraints to define a valid landmark detector

Related work:

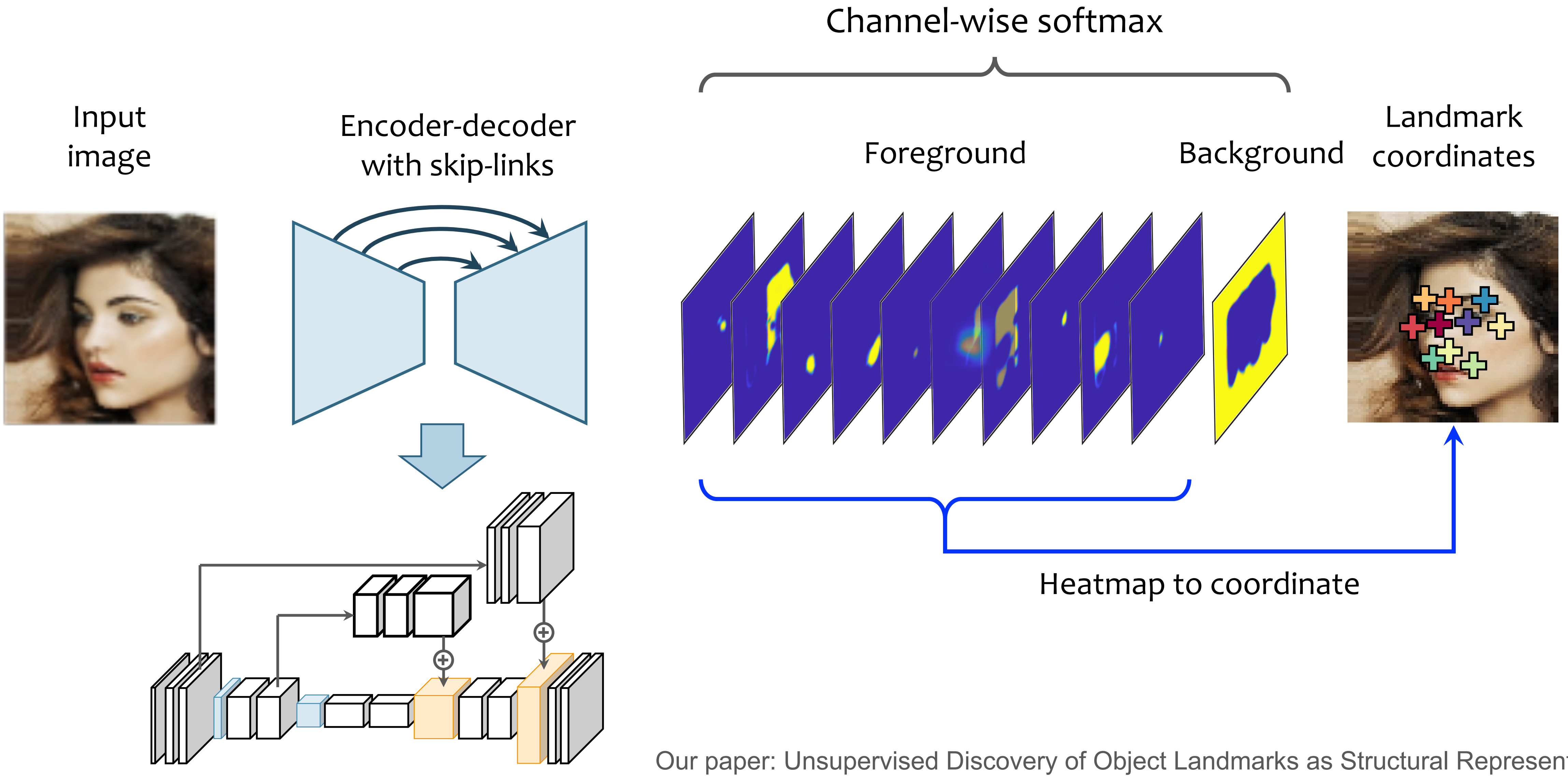
James Thewlis, Hakan Bilen, and Andrea Vedaldi,
"Unsupervised learning of object landmarks by factorized spatial embeddings,"
In *ICCV*, 2017.

Landmark detector: Architecture



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark detector: Architecture



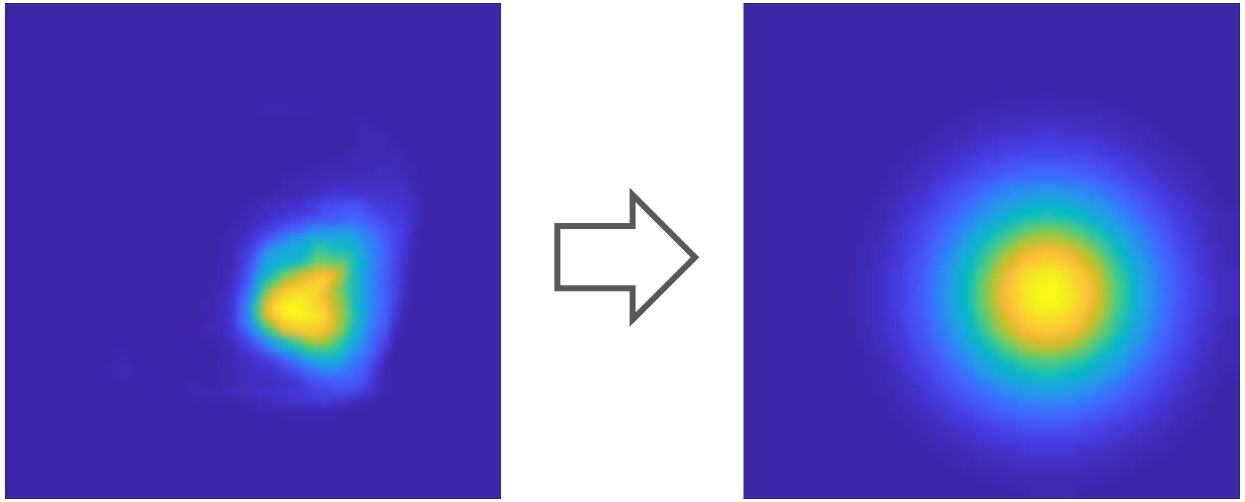
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

From heatmaps to coordinates

Ours:

A foreground heatmap

Isotropic Gaussian approximation



$$\mathcal{N}\left((x, y), \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}\right)$$

$\downarrow$

Landmark coordinate

- Averaged coordinate weighted by the heatmap
- (x, y) is **differentiable** with respect to the heatmap

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark discovery

$$\begin{rcases} (x_1,\ y_1) \\ (x_2,\ y_2) \\ \cdots \\ (x_K,\ y_K) \end{rcases}$$

Can be arbitrary without physical meanings

- The neural network **can be used** to output landmark coordinates.
- However, without additional training objectives, the landmark coordinates can be **arbitrary latent features**.

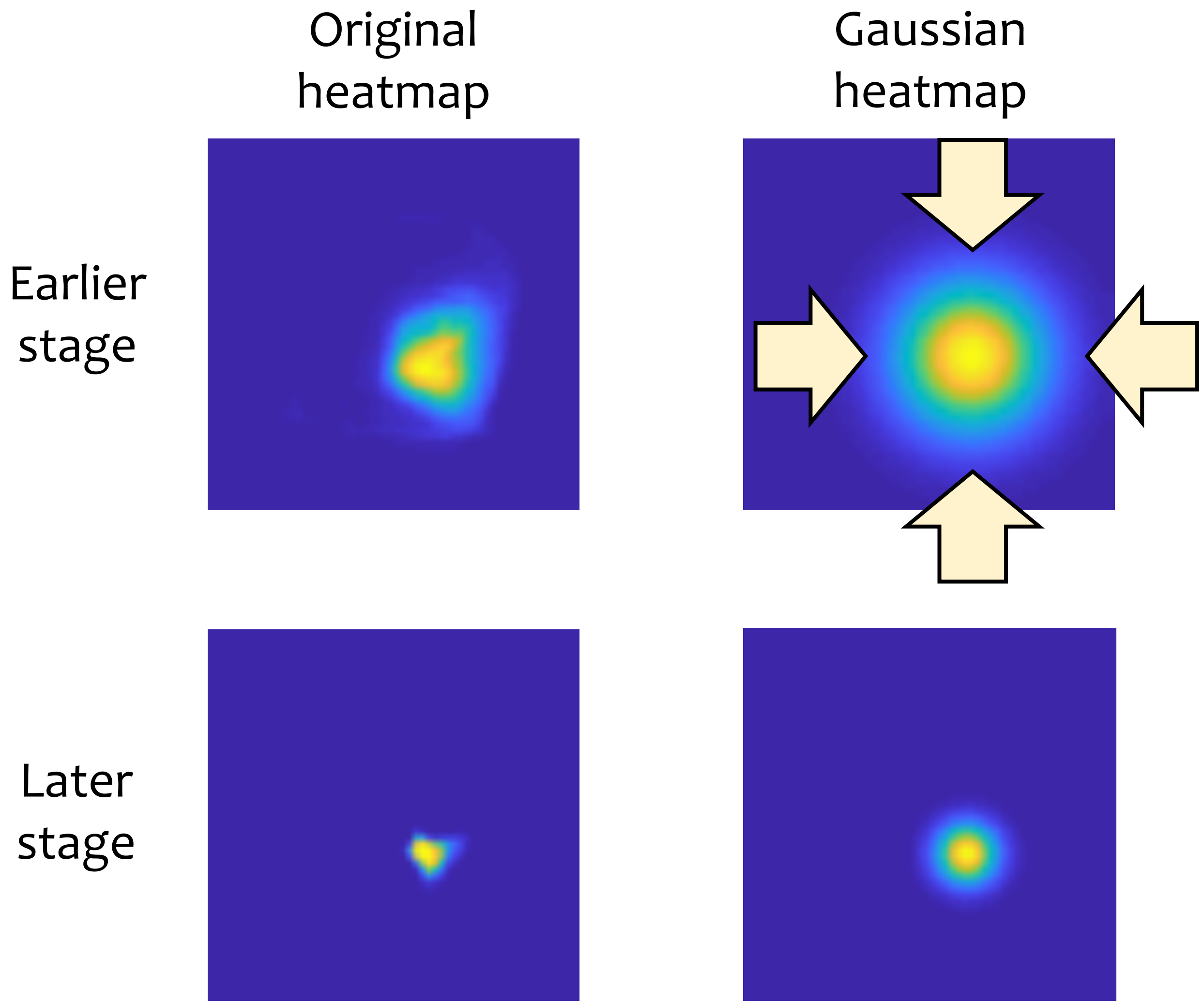
3 desirable properties for a landmark detector

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Property 1: **Concentration** of heatmap values

For a detector,
the output heatmap should
concentrate in a local region.

- Encourage the Gaussian variance to be small.



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

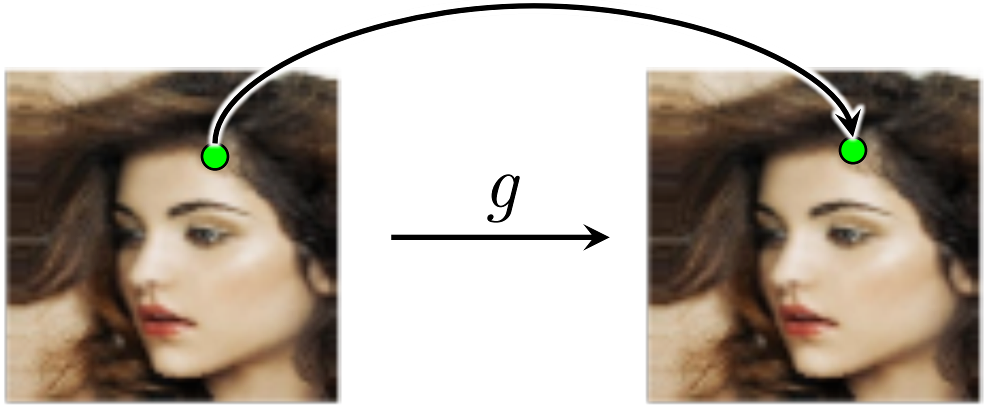
Property 2: **Separation** of landmarks

- Different landmarks should cover different visual semantics.
- Penalize if the **pairwise distances** among landmarks are too small.

$$L_{\text{sep}} = \sum_{k \neq k'}^{1,\ldots,K} \exp\left(- \frac{\|(x_{k'}, y_{k'}) - (x_k, y_k)\|_2^2}{2\sigma_{\text{sep}}^2} \right)$$

Property 3: **Equivariance**

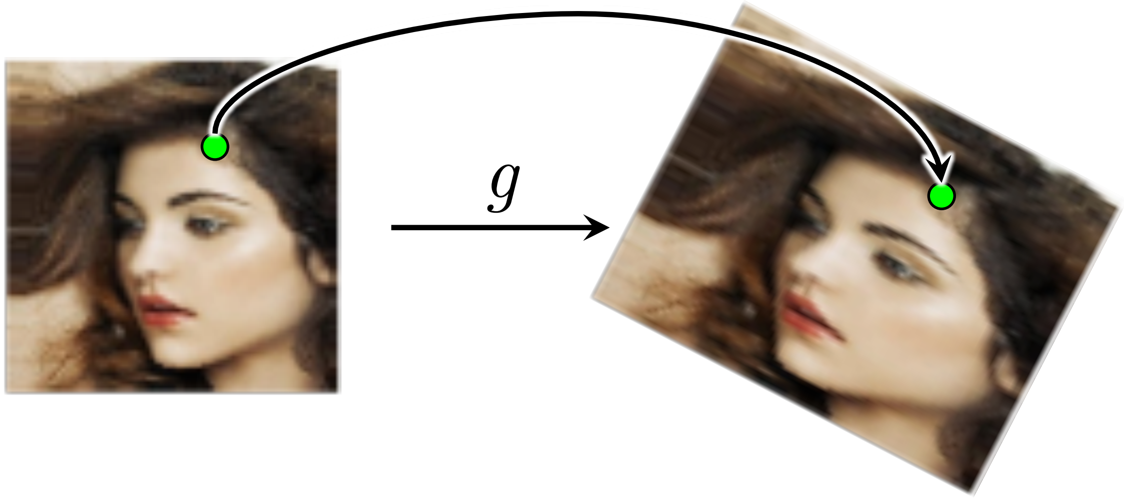
- For a transformation g that does not change local visual semantics.
- The landmarks on the two images should satisfy the same transformation g.



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

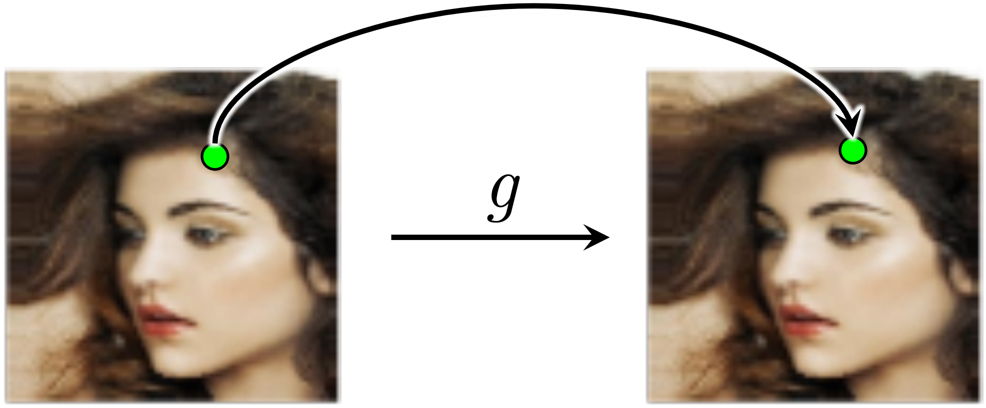
Property 3: **Equivariance**

- For a transformation g that does not change local visual semantics.
- The landmarks on the two images should satisfy the same transformation g.



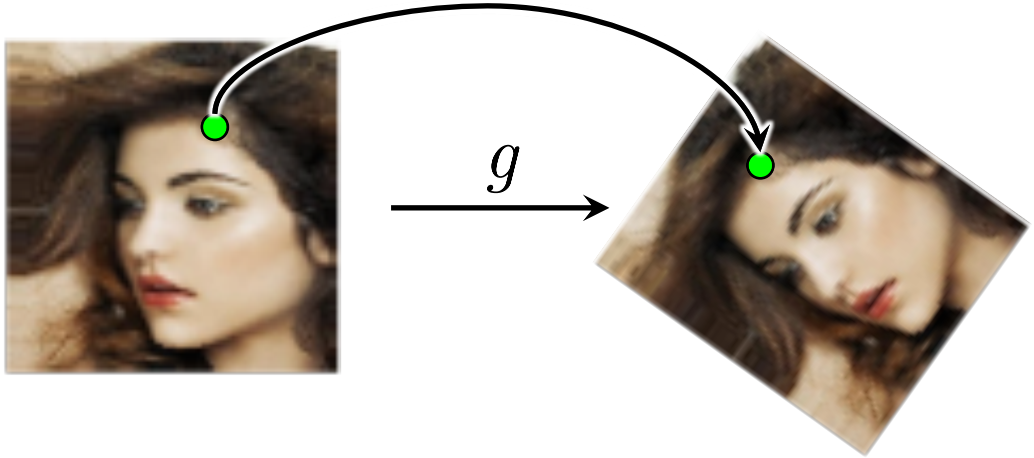
Property 3: **Equivariance**

- For a transformation g that does not change local visual semantics.
- The landmarks on the two images should satisfy the same transformation g.



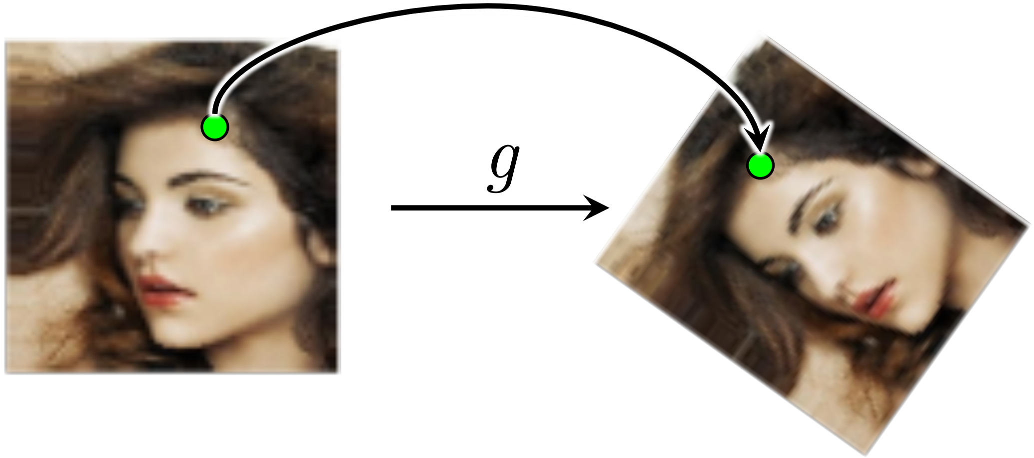
Property 3: **Equivariance**

- For a transformation g that does not change local visual semantics.
- The landmarks on the two images should satisfy the same transformation g.



Property 3: **Equivariance**

- For a transformation g that does not change local visual semantics.
- The landmarks on the two images should satisfy the same transformation g.



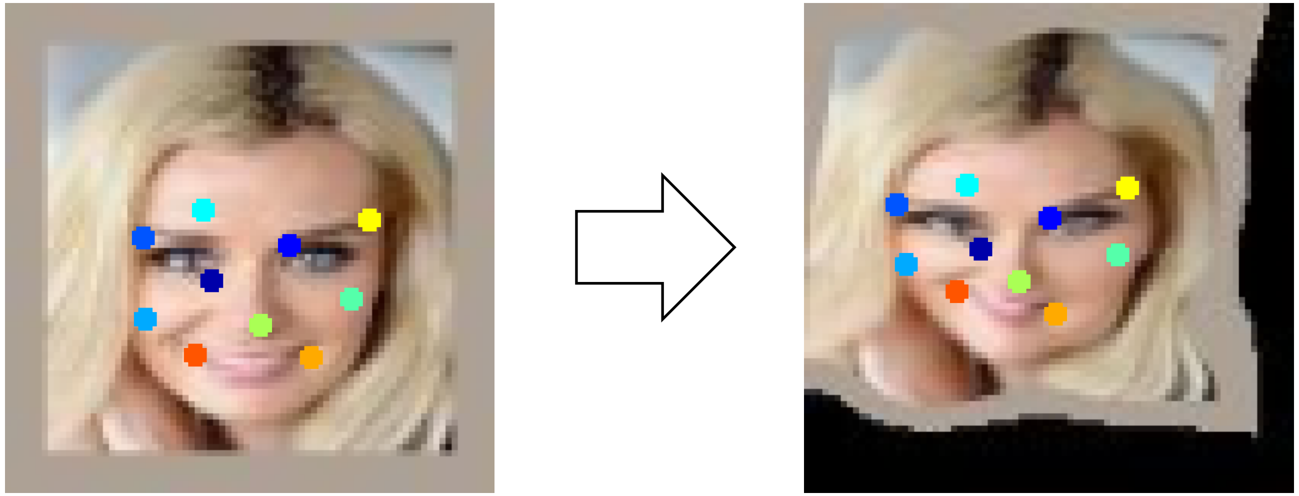
$$L_{\mathrm{eqv}} = \sum_{k=1}^{K} \left\| g(x'_k, y'_k) - (x_k, y_k) \right\|_2^2$$

- Equivariance for landmark discovery has been explored by Thewlis et al, 2017.
- Ours are directly formulated on the landmark coordinate.

(Thewlis et al, 2017) James Thewlis, Hakan Bilen, and Andrea Vedaldi, "Unsupervised learning of object landmarks by factorized spatial embeddings," In *ICCV*, 2017.

Property 3: **Equivariance** – the transformation

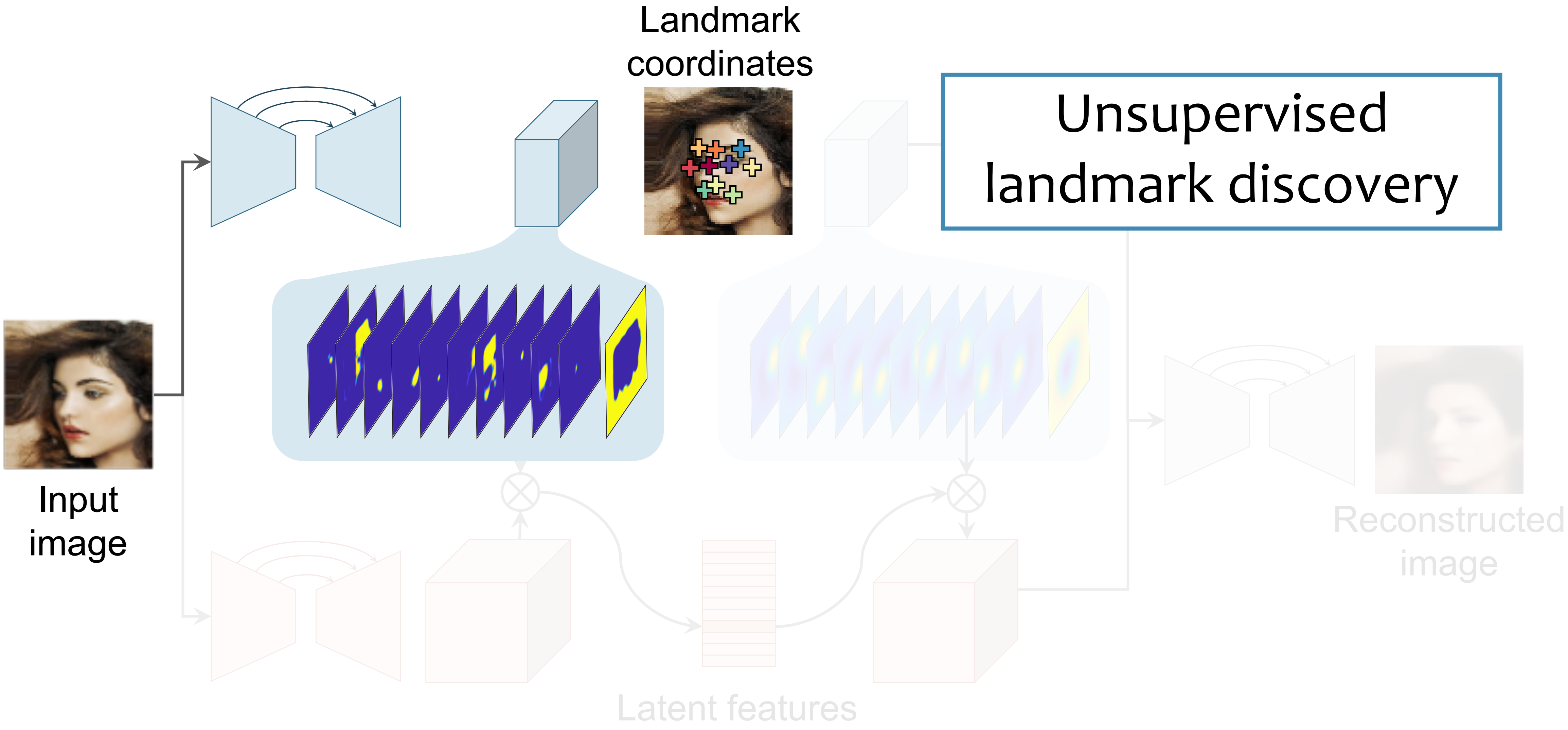
- Random thin-plate-spline (TPS) to synthesize the transformation g
  - Global affine: Translation, Scaling, Rotation
  - Local TPS:



- For videos, also use the optical flows as the transformation g

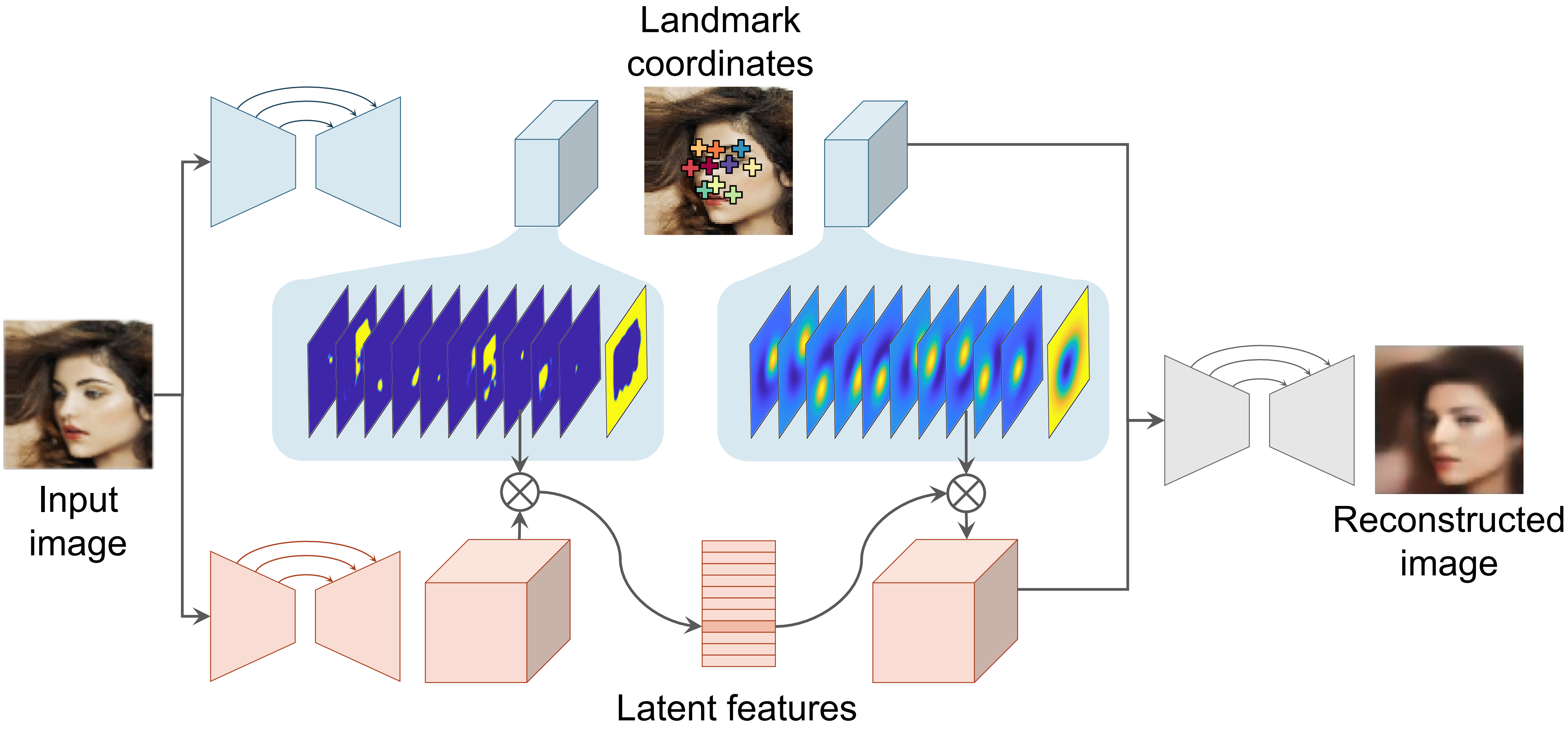
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture



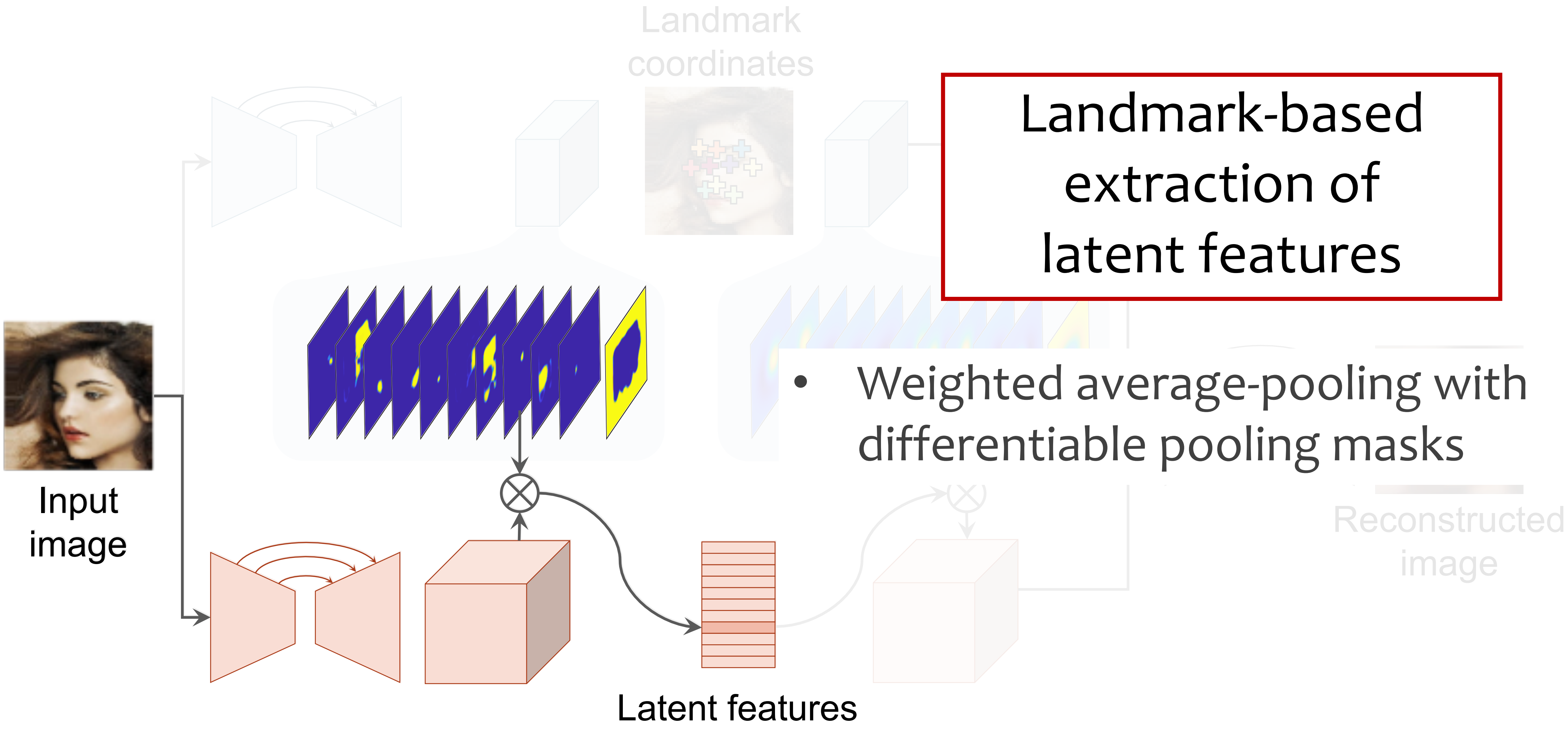
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture



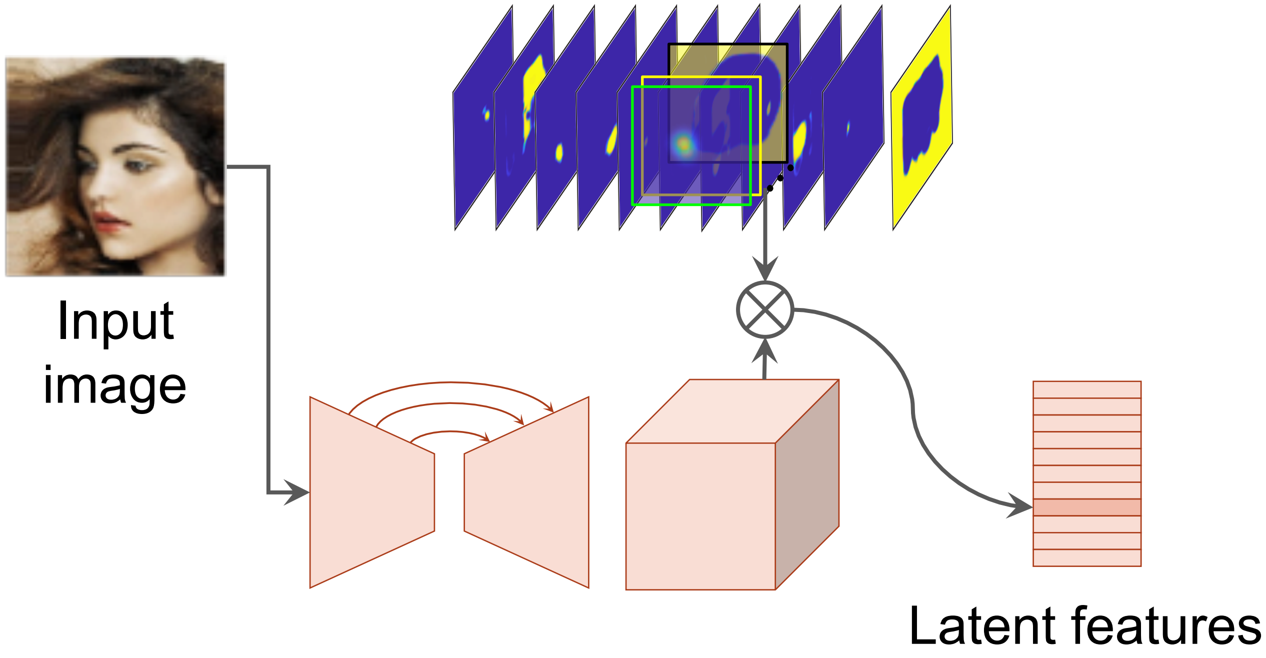
Landmark-based extraction of latent features

- Weighted average-pooling with differentiable pooling masks

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

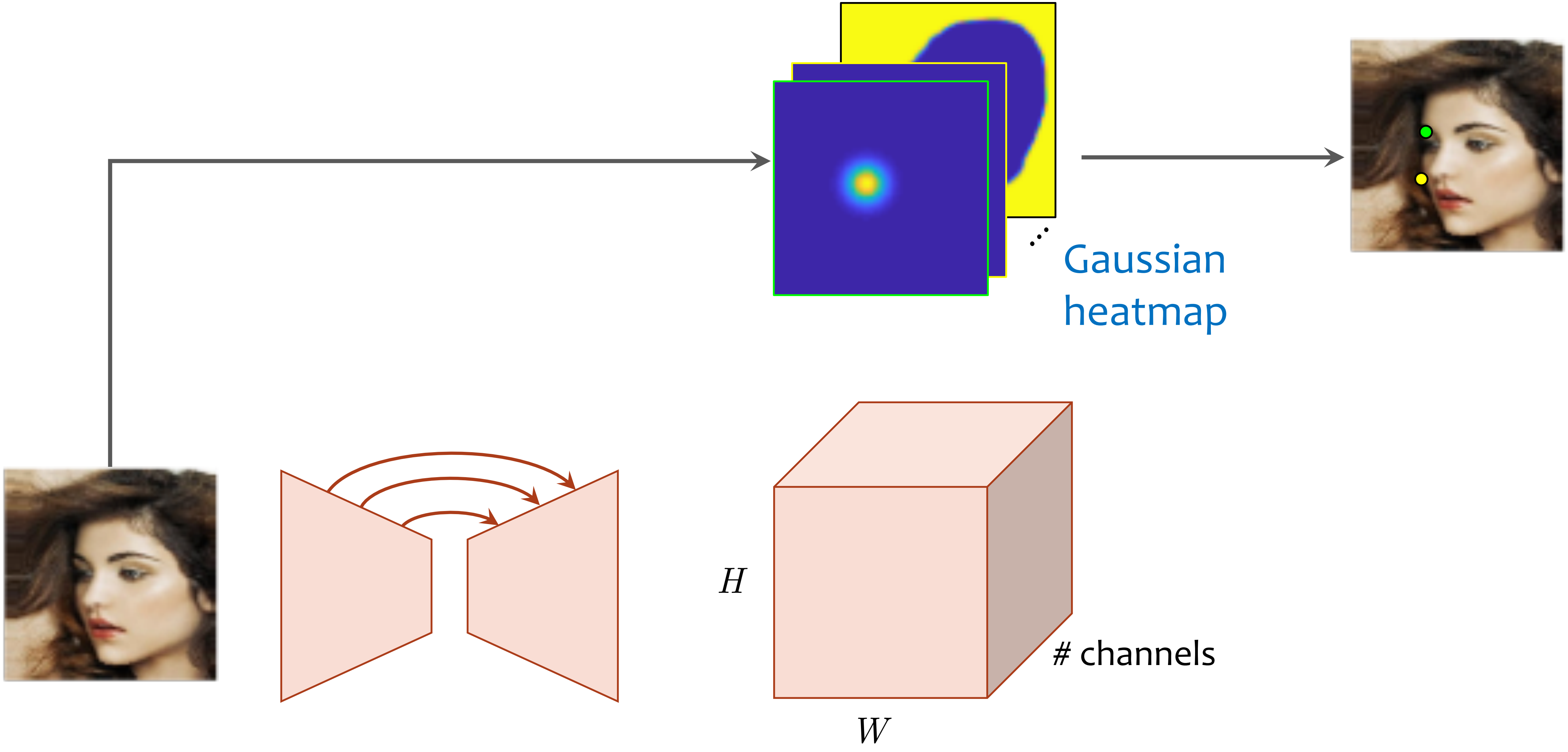
Overview of our neural network architecture

Landmark-based extraction of latent features



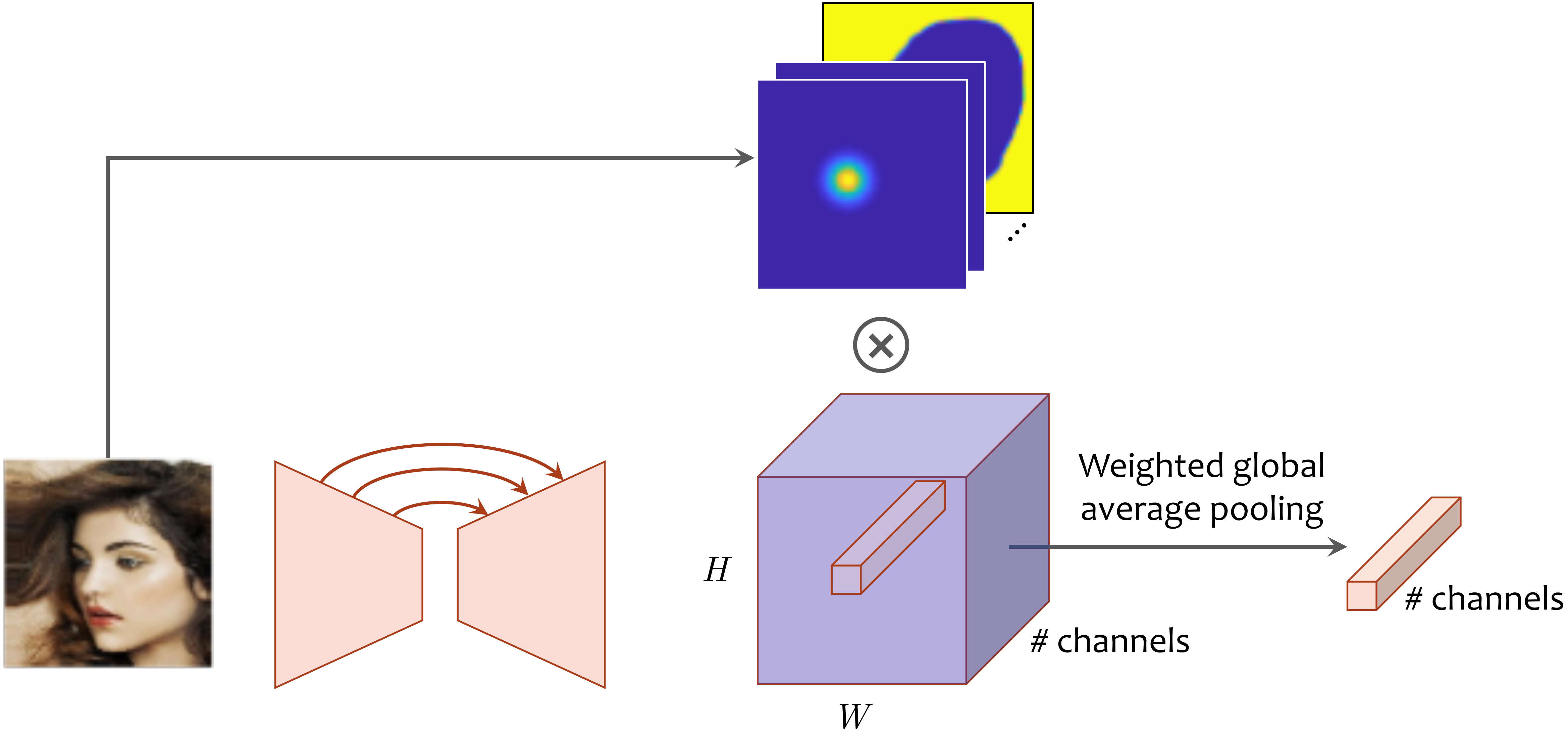
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark-based feature extraction



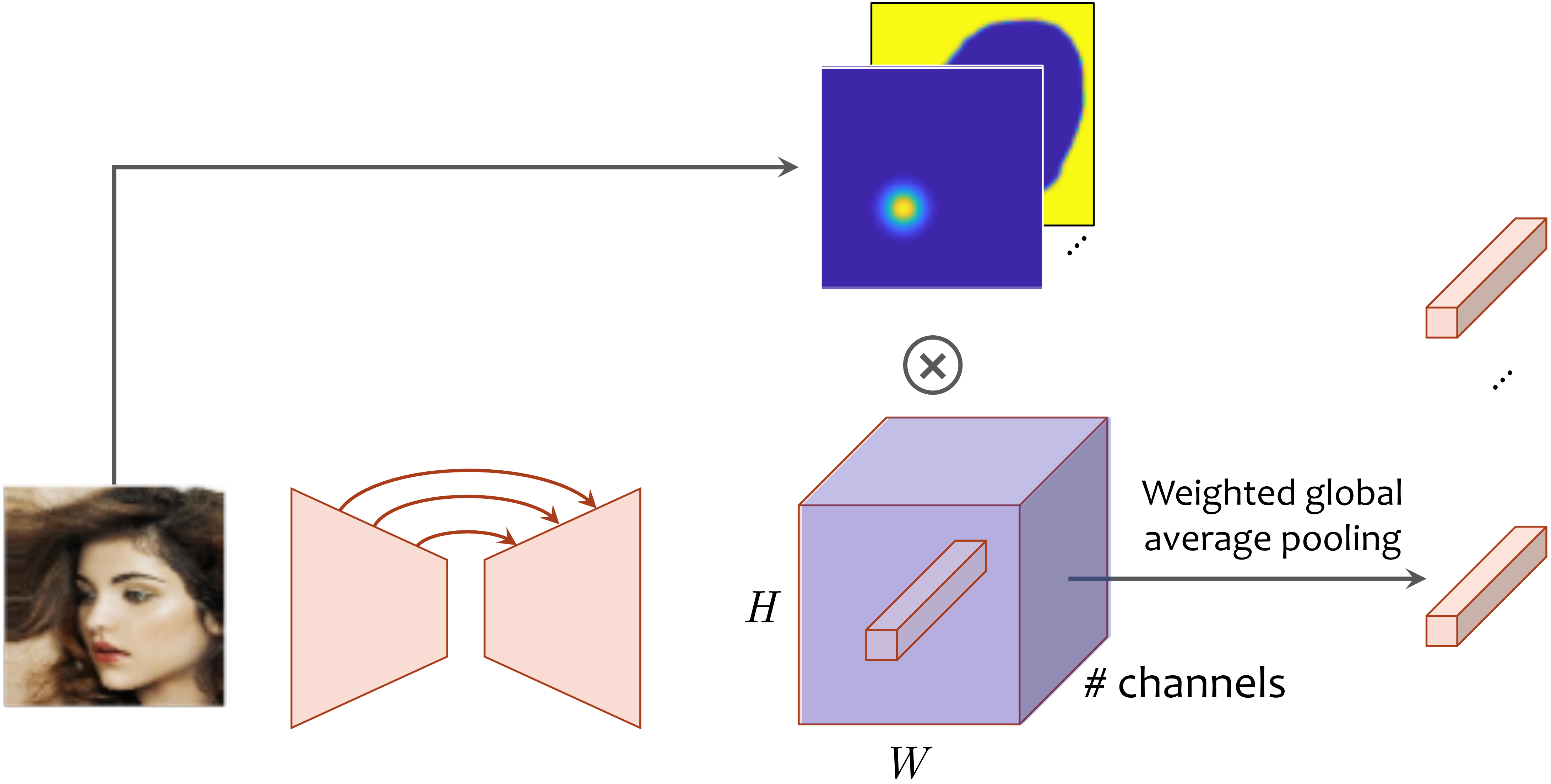
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark-based feature extraction



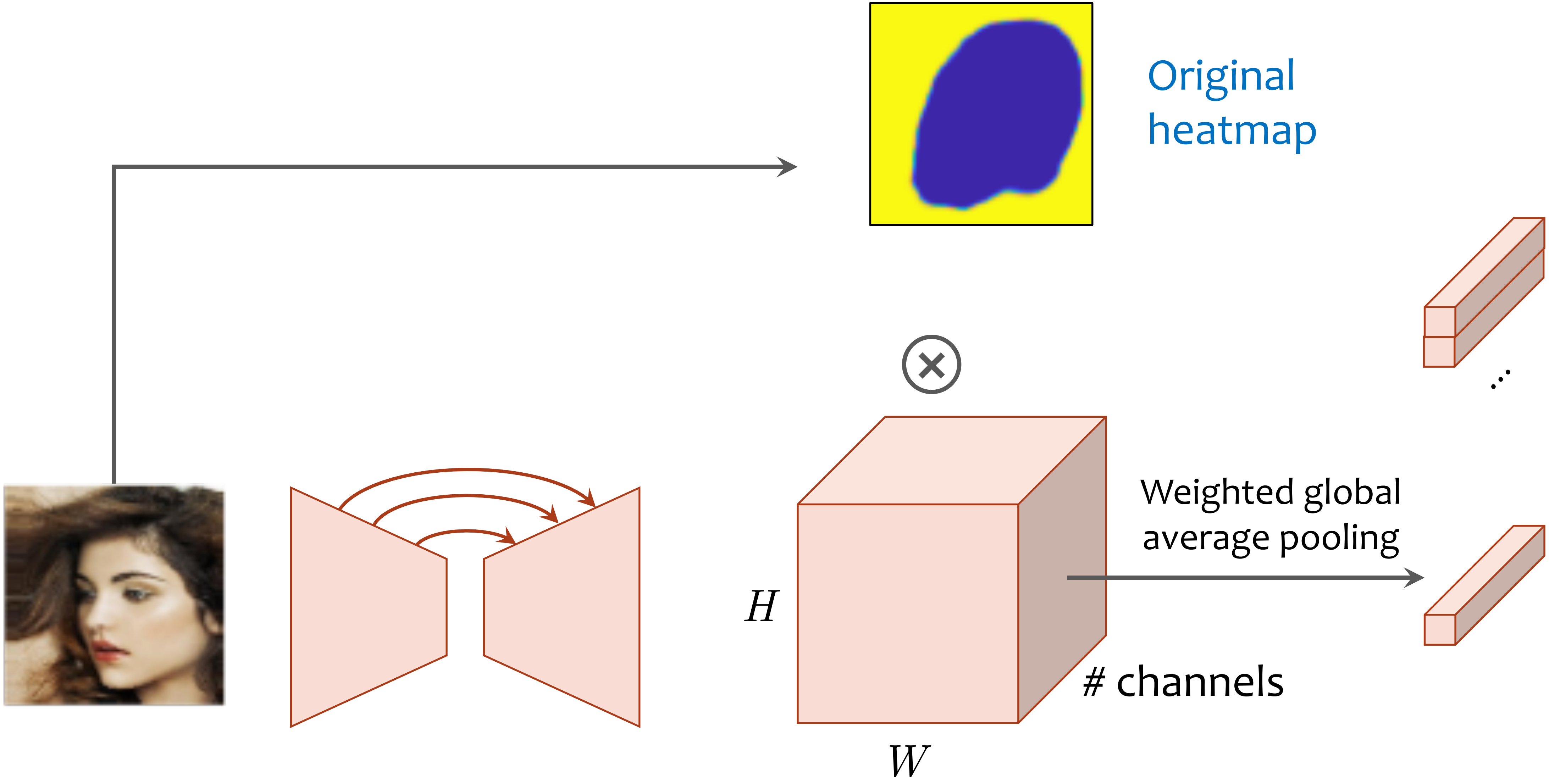
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark-based feature extraction



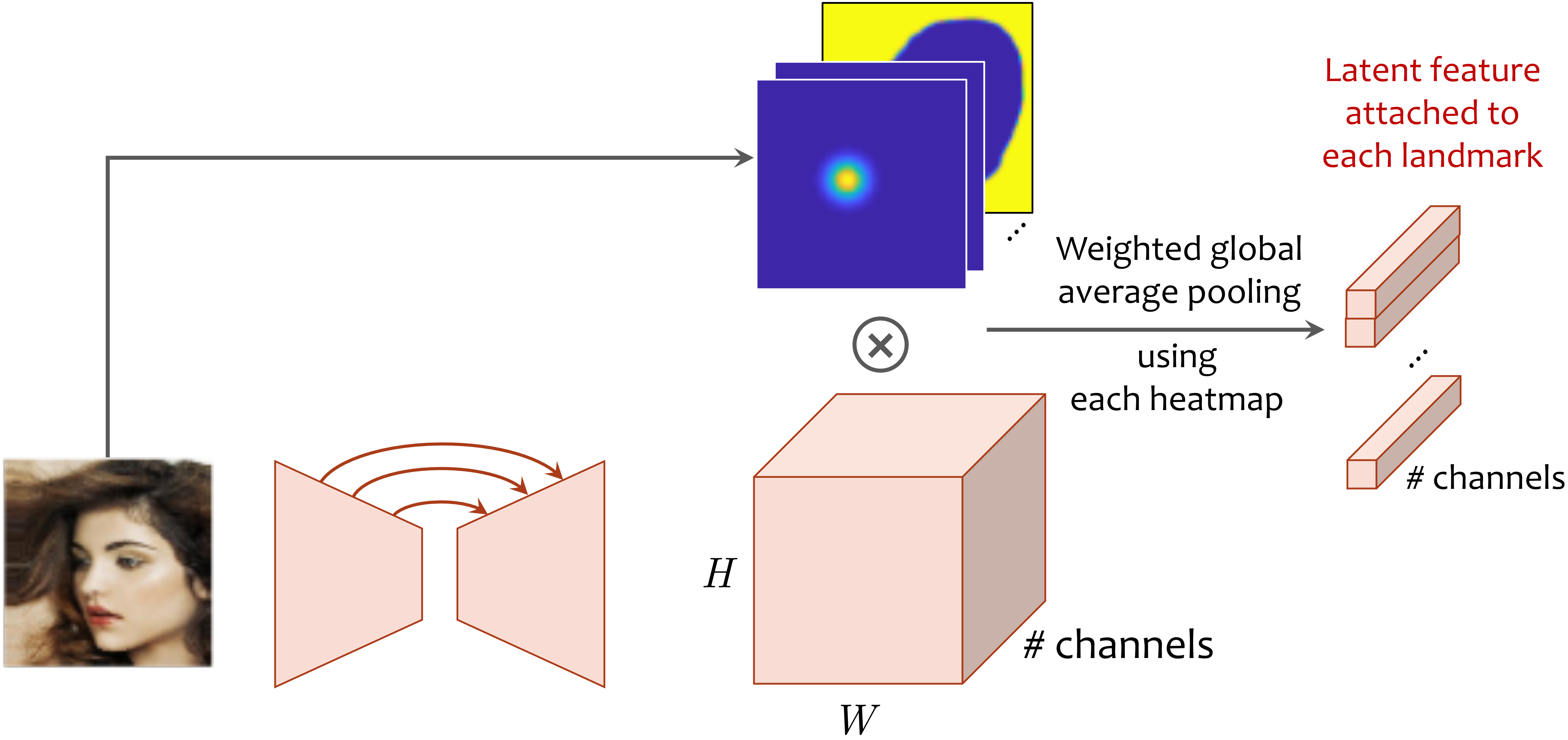
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark-based feature extraction

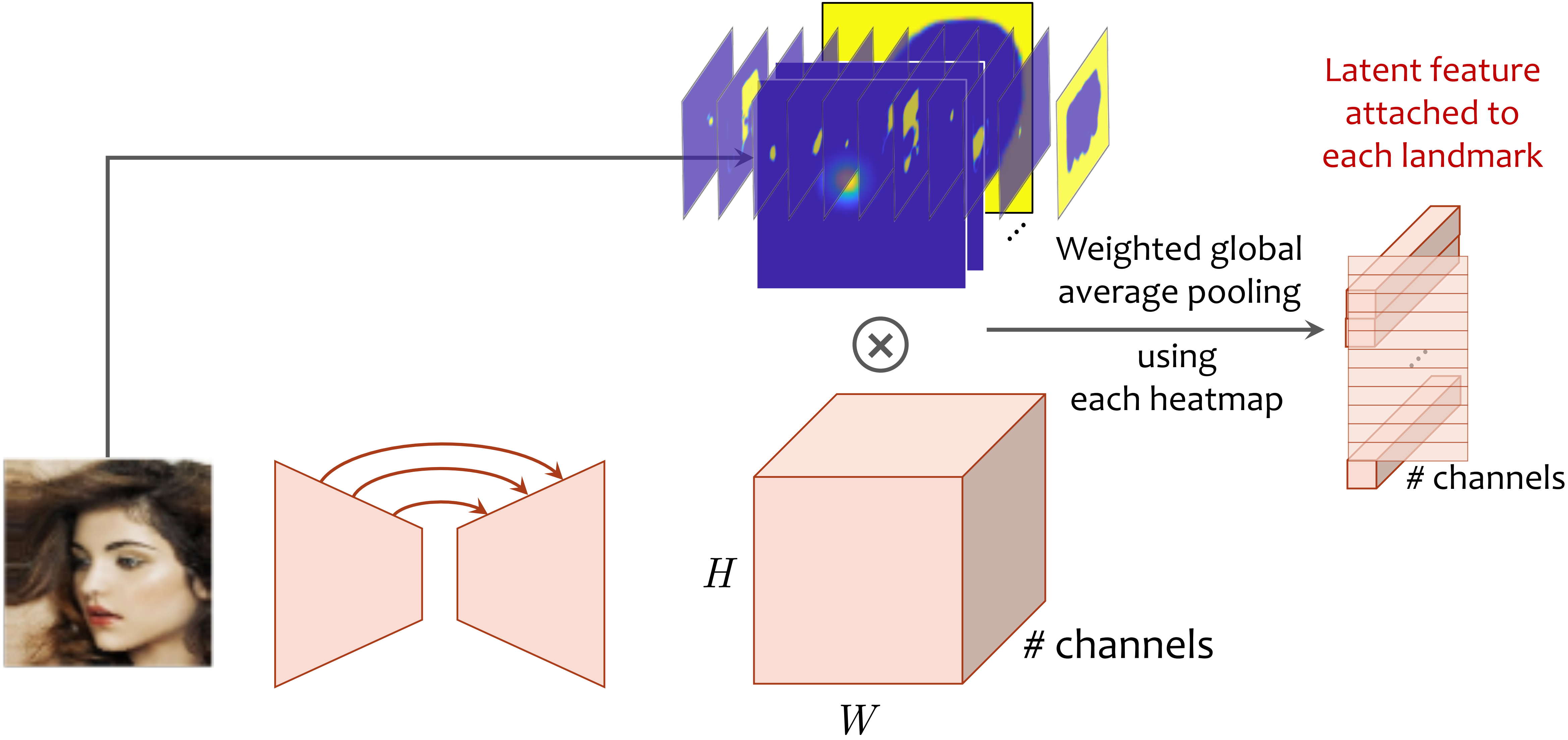


Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark-based feature extraction



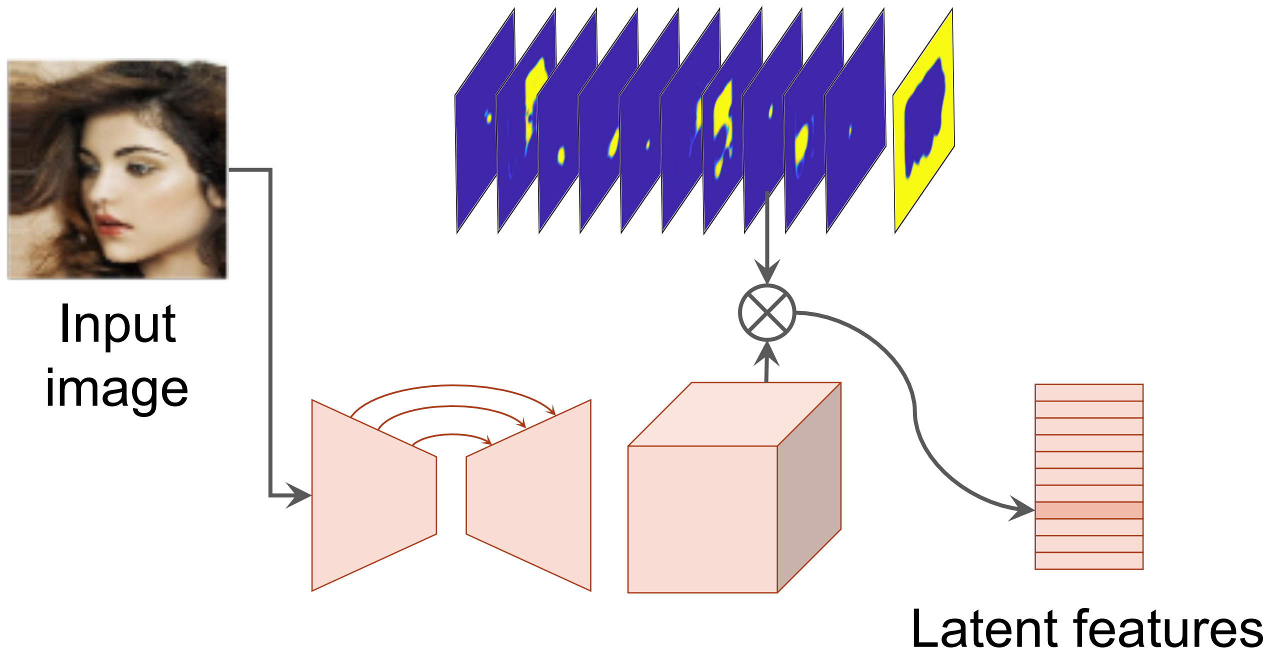
Landmark-based feature extraction



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

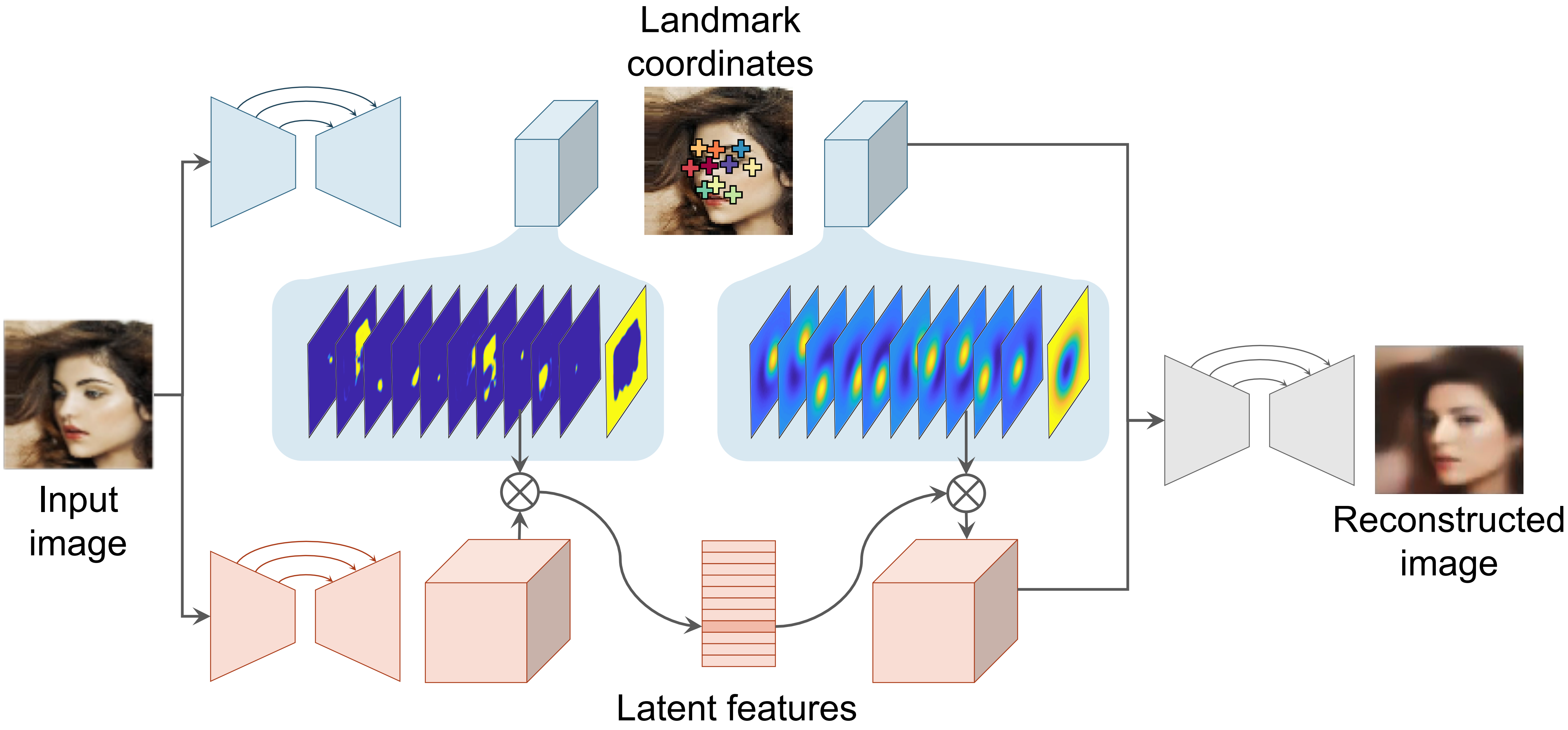
Overview of our neural network architecture

Landmark-based extraction of latent features



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture

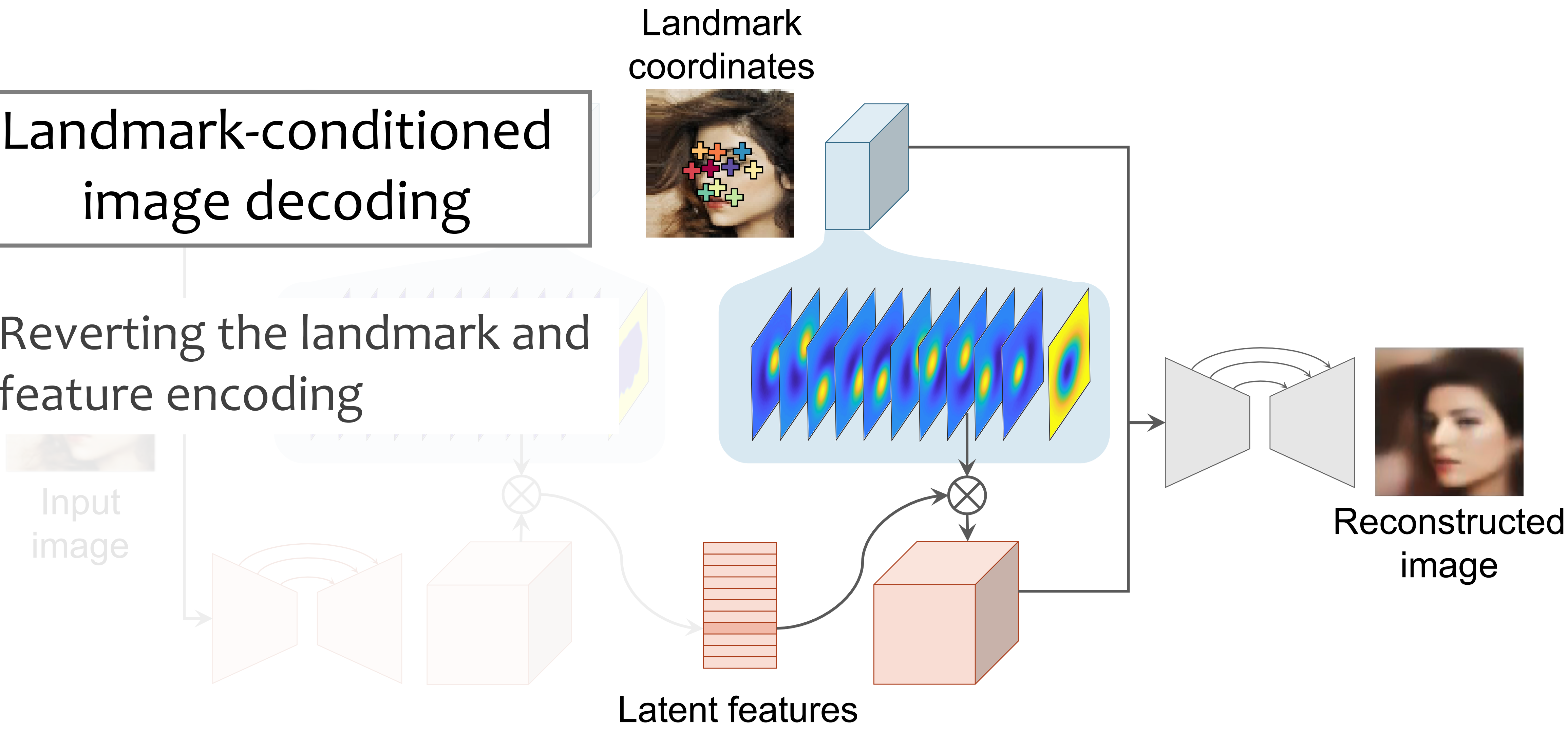


Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture

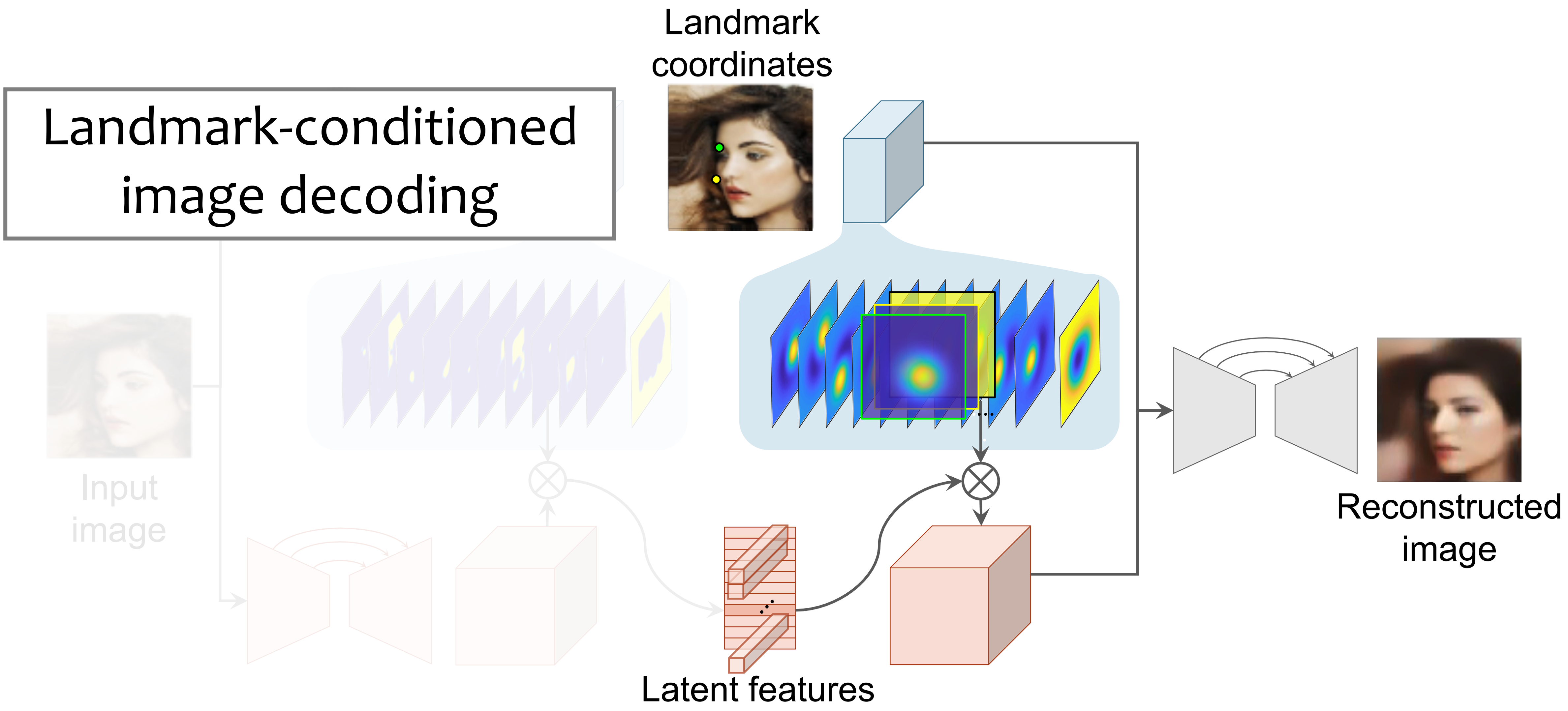
Landmark-conditioned image decoding

- Reverting the landmark and feature encoding



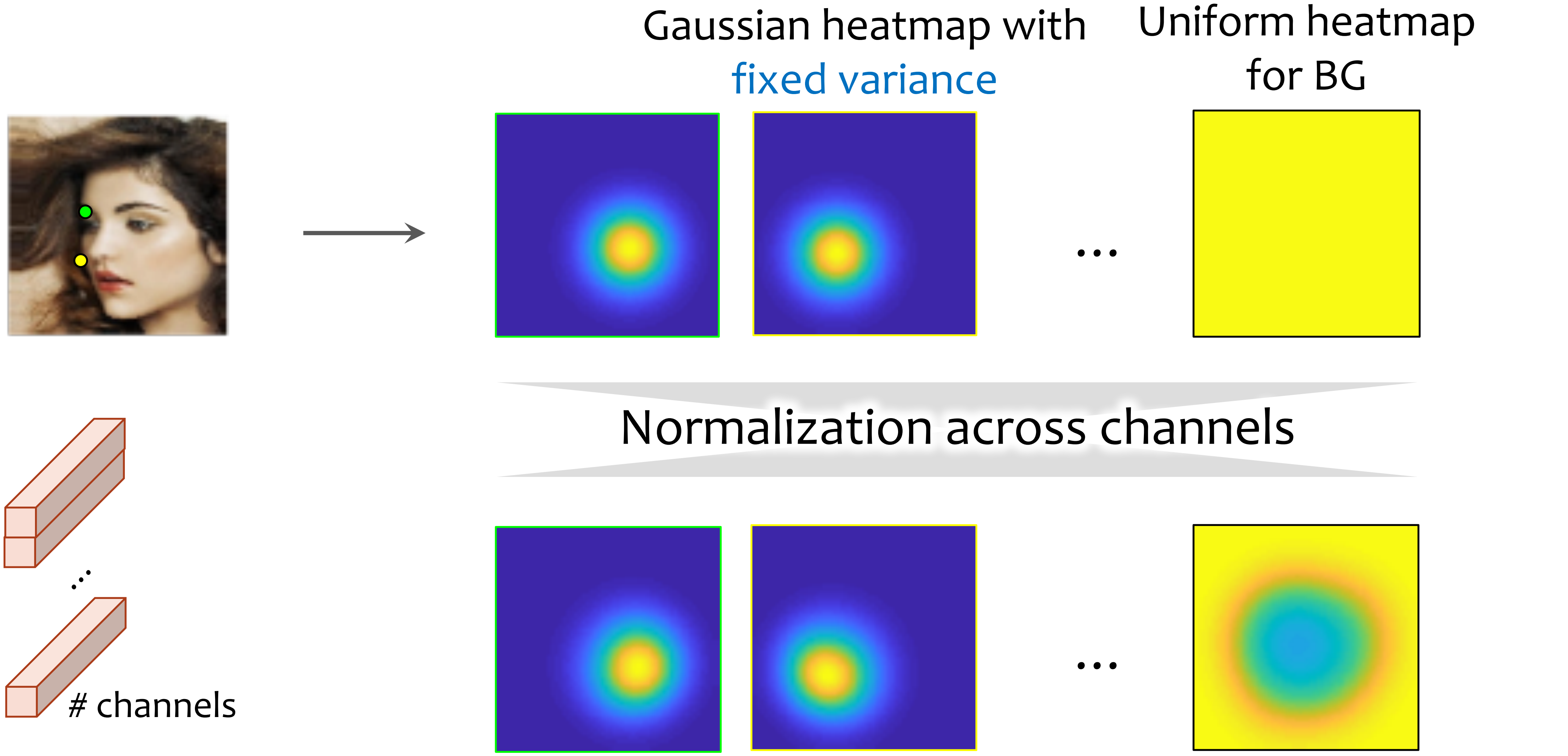
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture



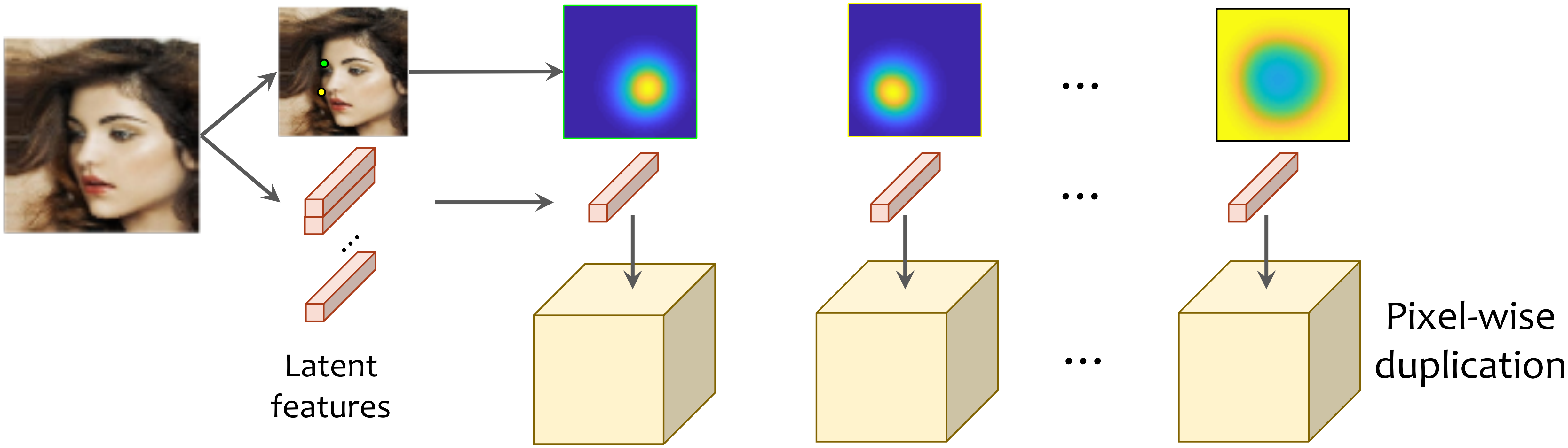
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Landmark-based decoding

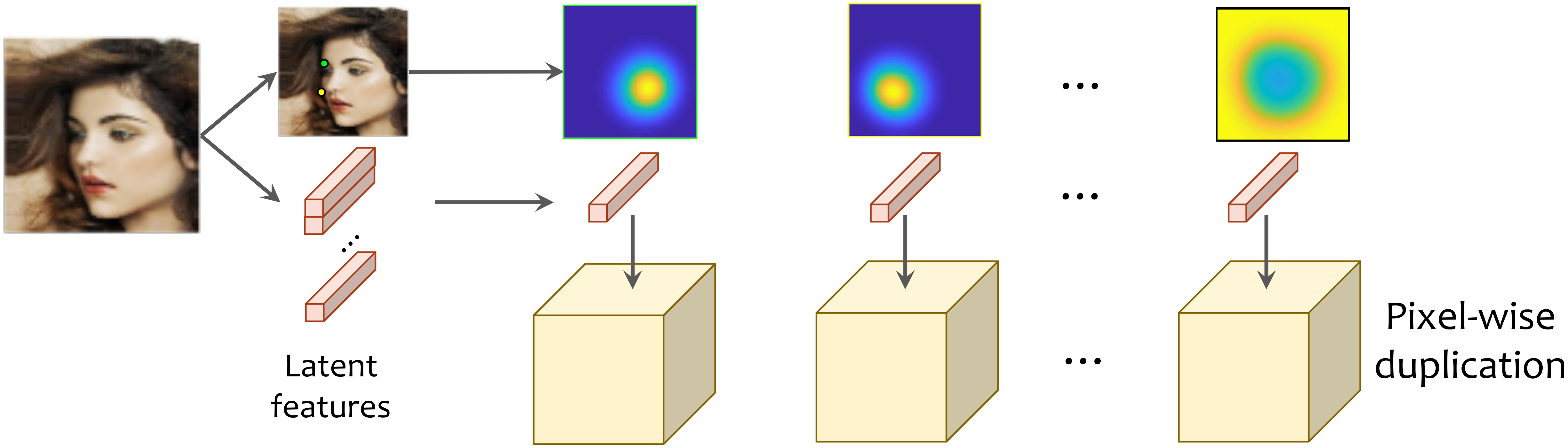


Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

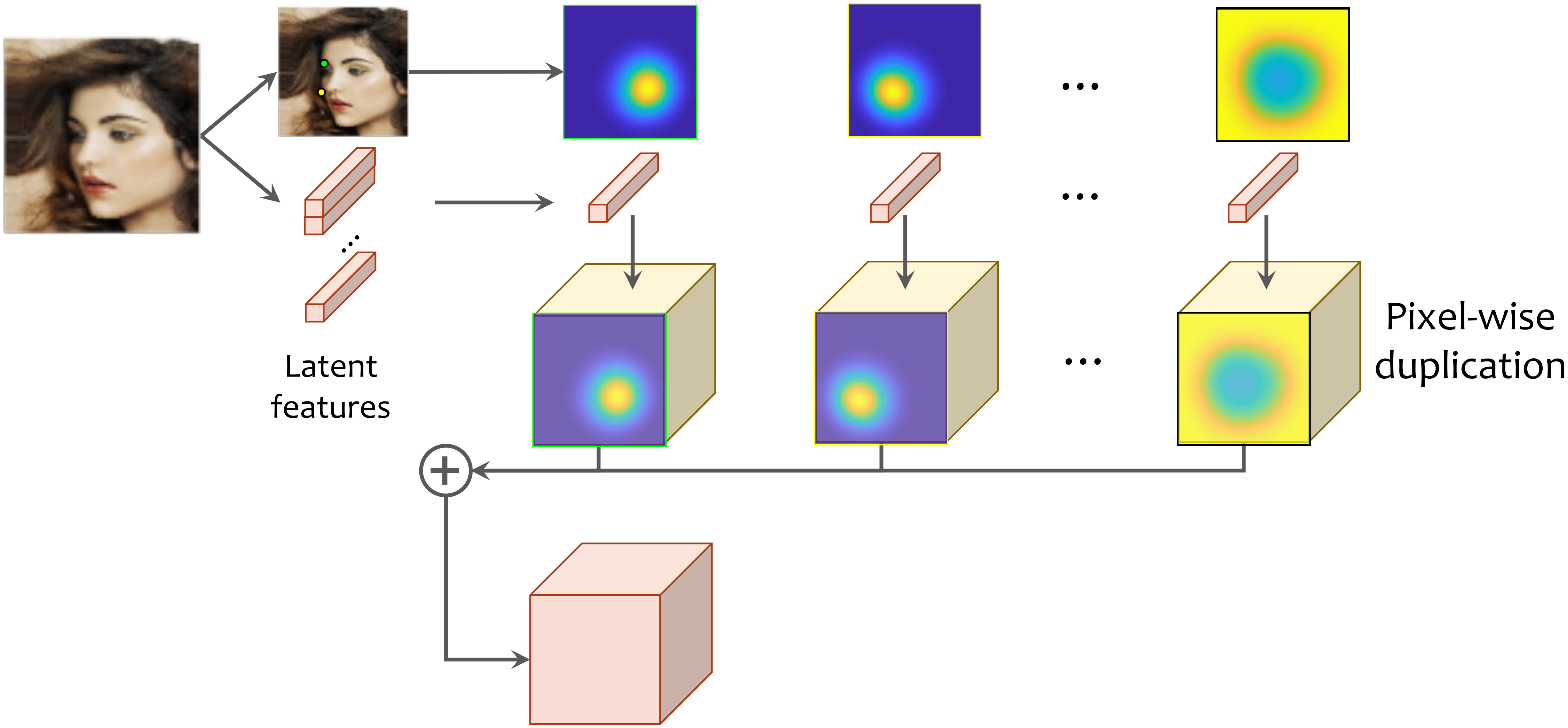
Landmark-based decoding: all landmarks



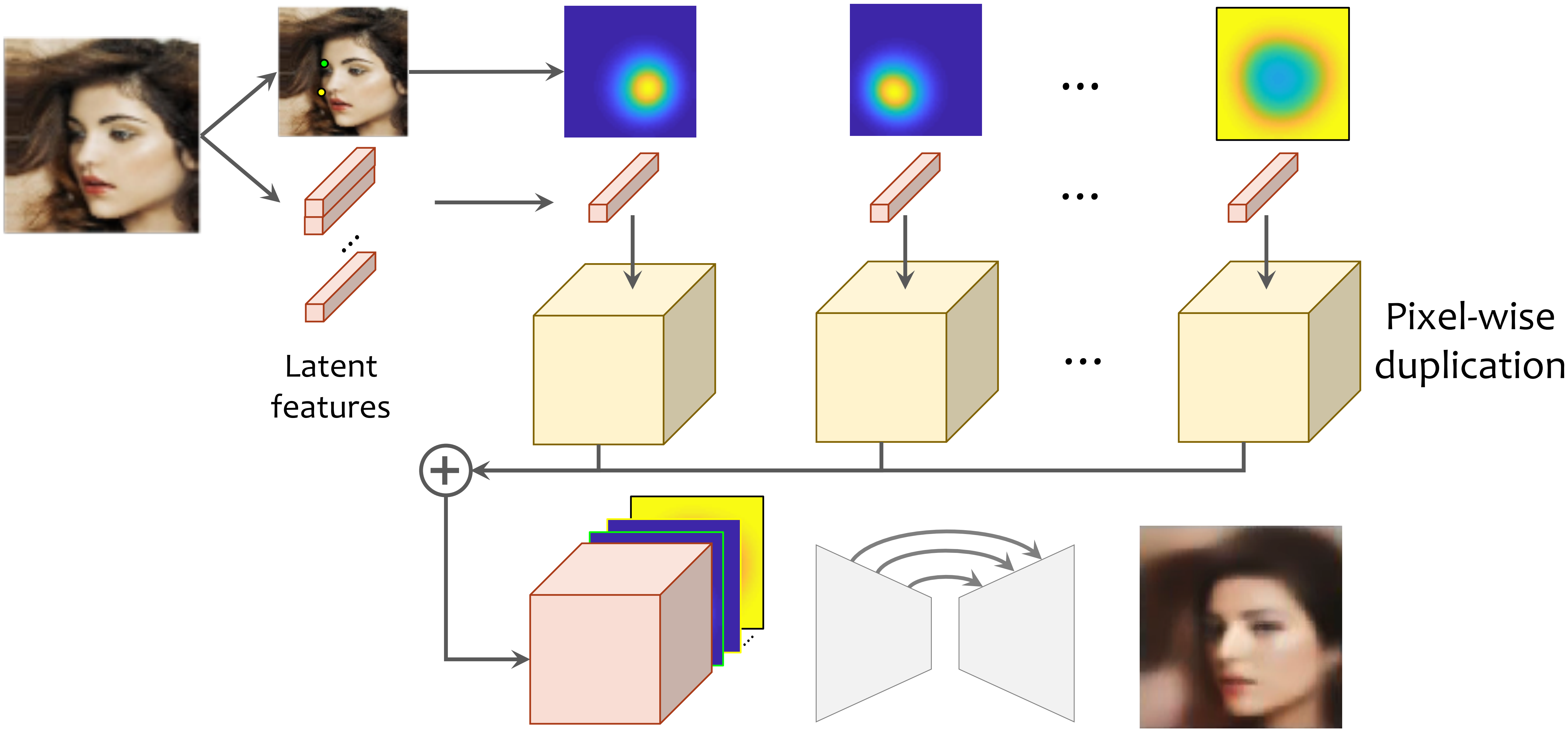
Landmark-based decoding: all landmarks



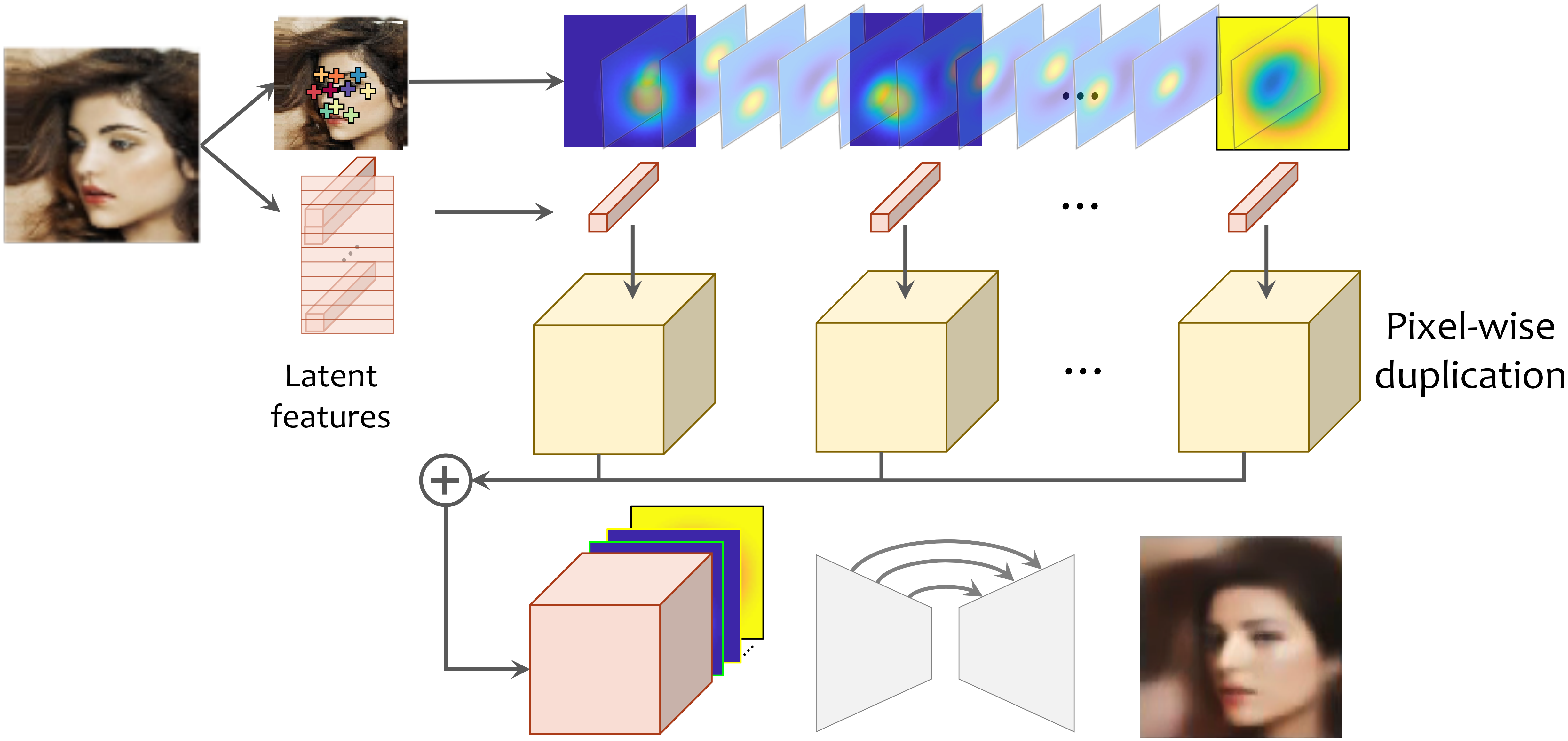
Landmark-based decoding: all landmarks



Landmark-based decoding: all landmarks

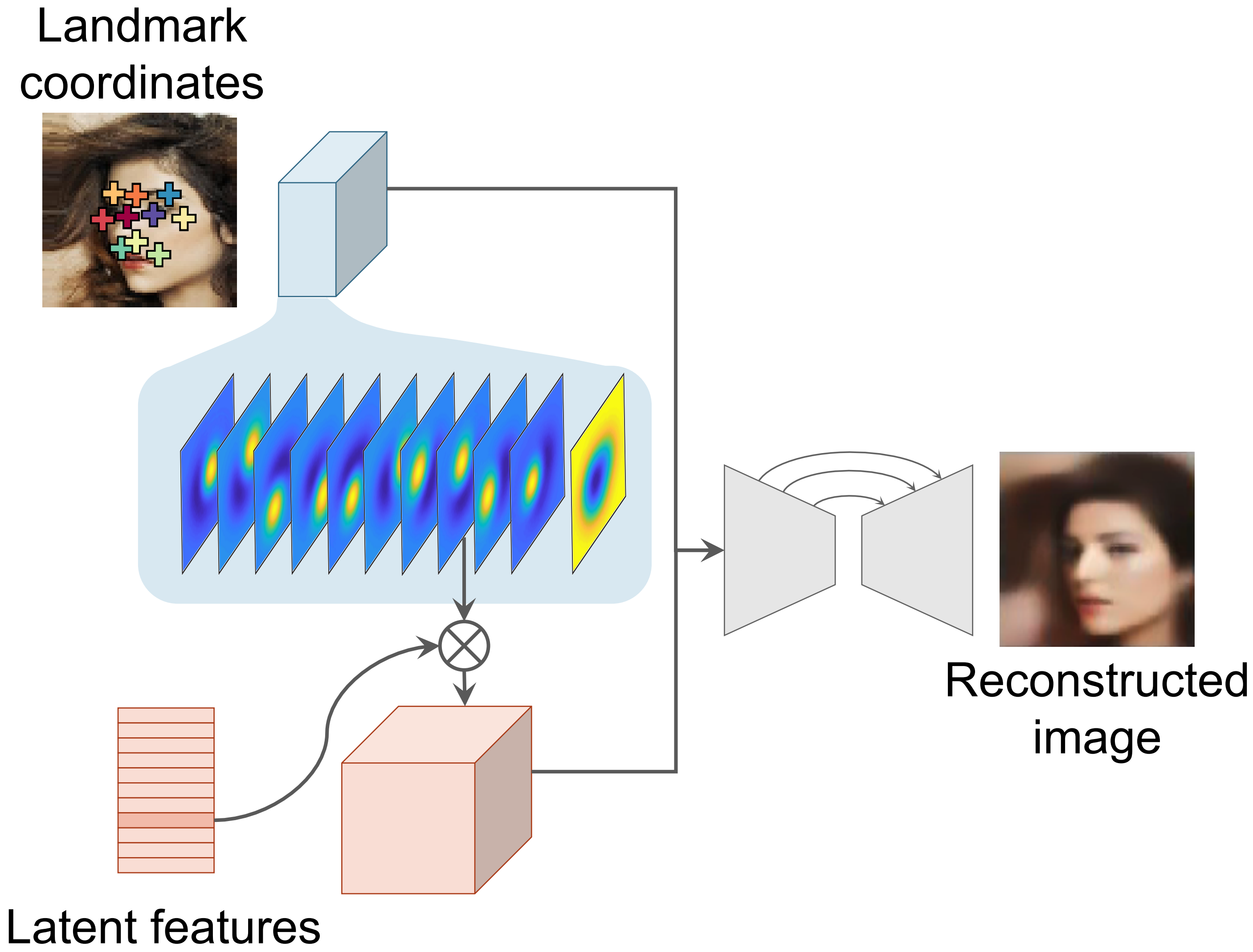


Landmark-based decoding: all landmarks



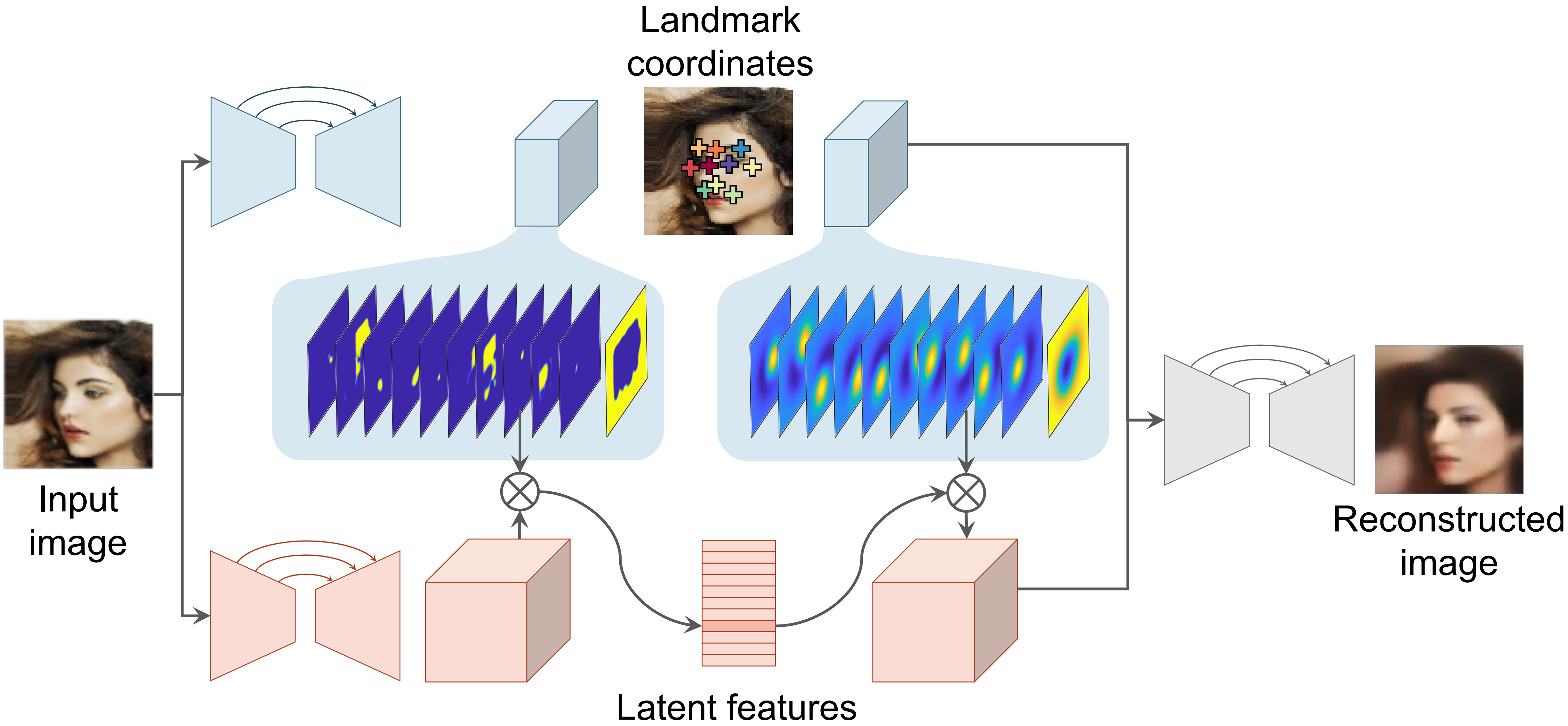
Overview of our neural network architecture

Landmark-conditioned image decoding



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Overview of our neural network architecture

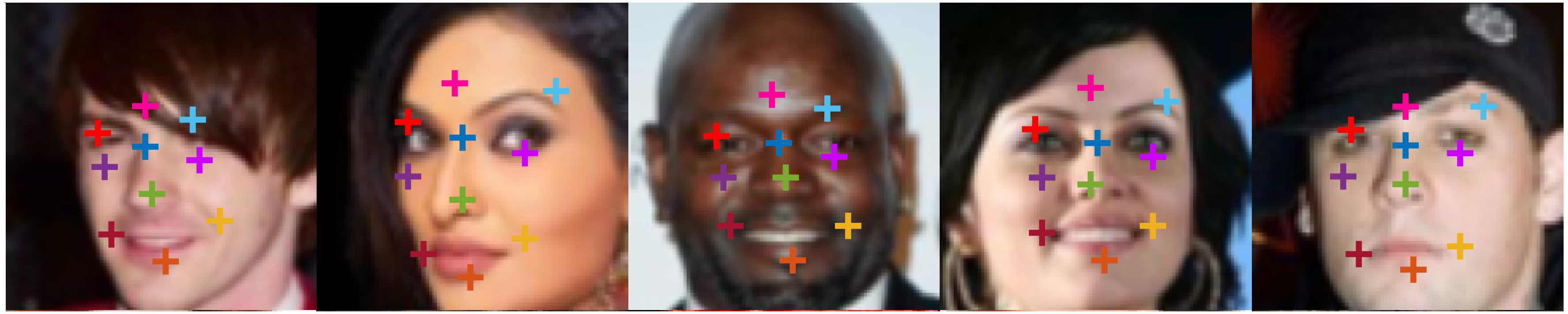


Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

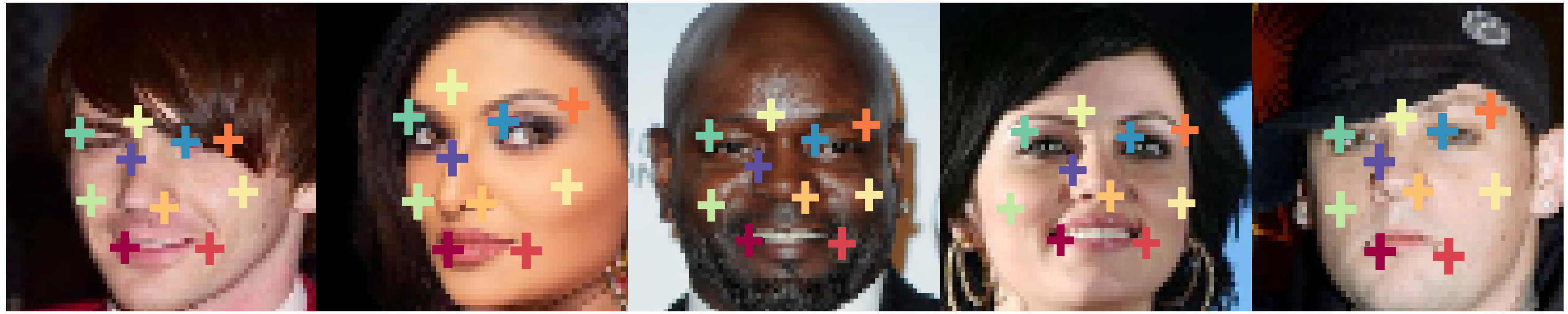
Experimental results

Landmark discovery: Faces, 10 landmarks

Thewlis at al.



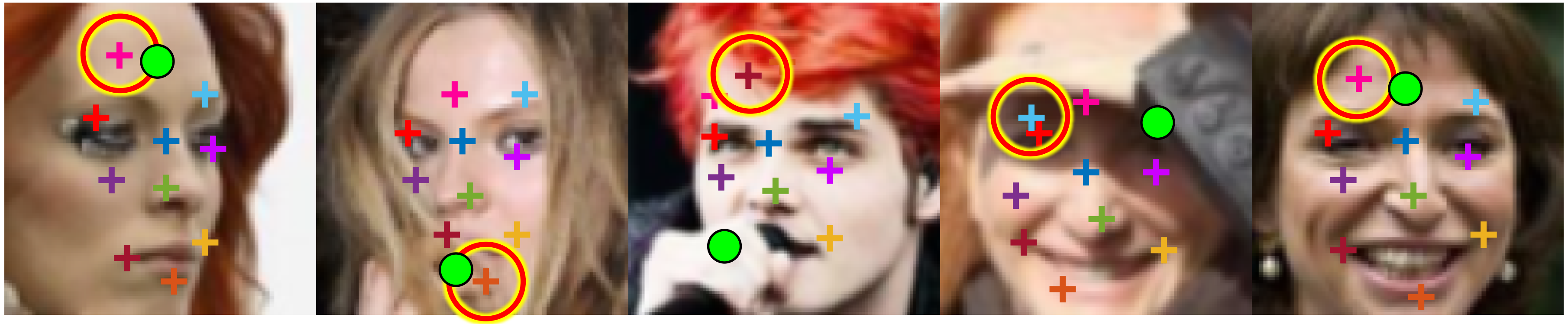
Ours



(Thewlis et al.) James Thewlis, Hakan Bilen, and Andrea Vedaldi, "Unsupervised learning of object landmarks by factorized spatial embeddings," In *ICCV*, 2017.

Unsupervised discovery: Faces, 10 landmarks

Thewlis at al.



Errors	Forehead landmark to the left	Lower-lip landmark to the right	Mouth-corner landmark on the forehead	Right-eyebrow landmark on the left side	Forehead landmark to the left

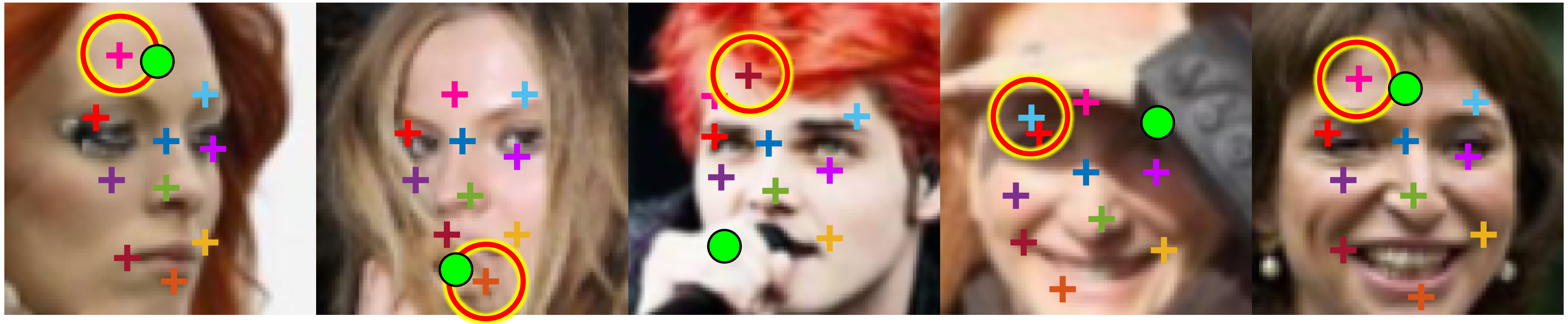
Incorrect landmarks

Expected correct location

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised discovery: Faces, 10 landmarks

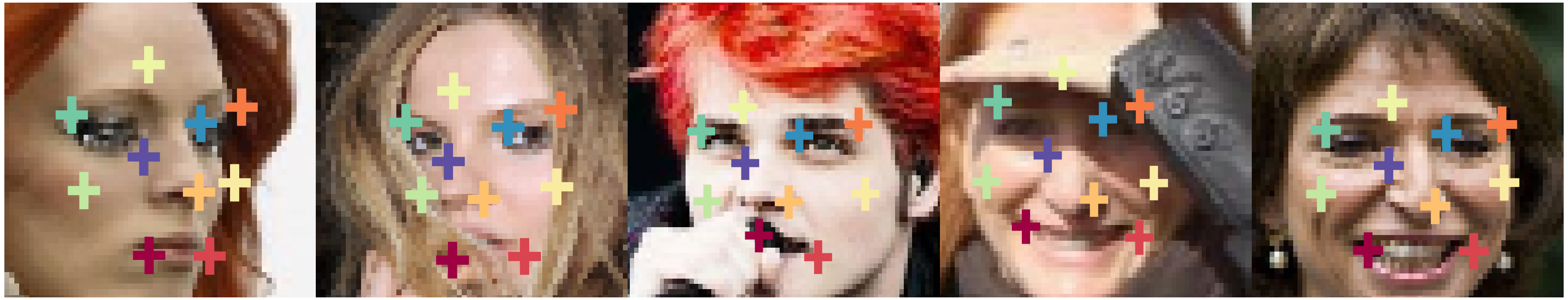
Thewlis at al.



Errors

Forehead landmark to the left	Lower-lip landmark to the right	Mouth-corner landmark on the forehead	Right-eyebrow landmark on the left side	Forehead landmark to the left

Ours

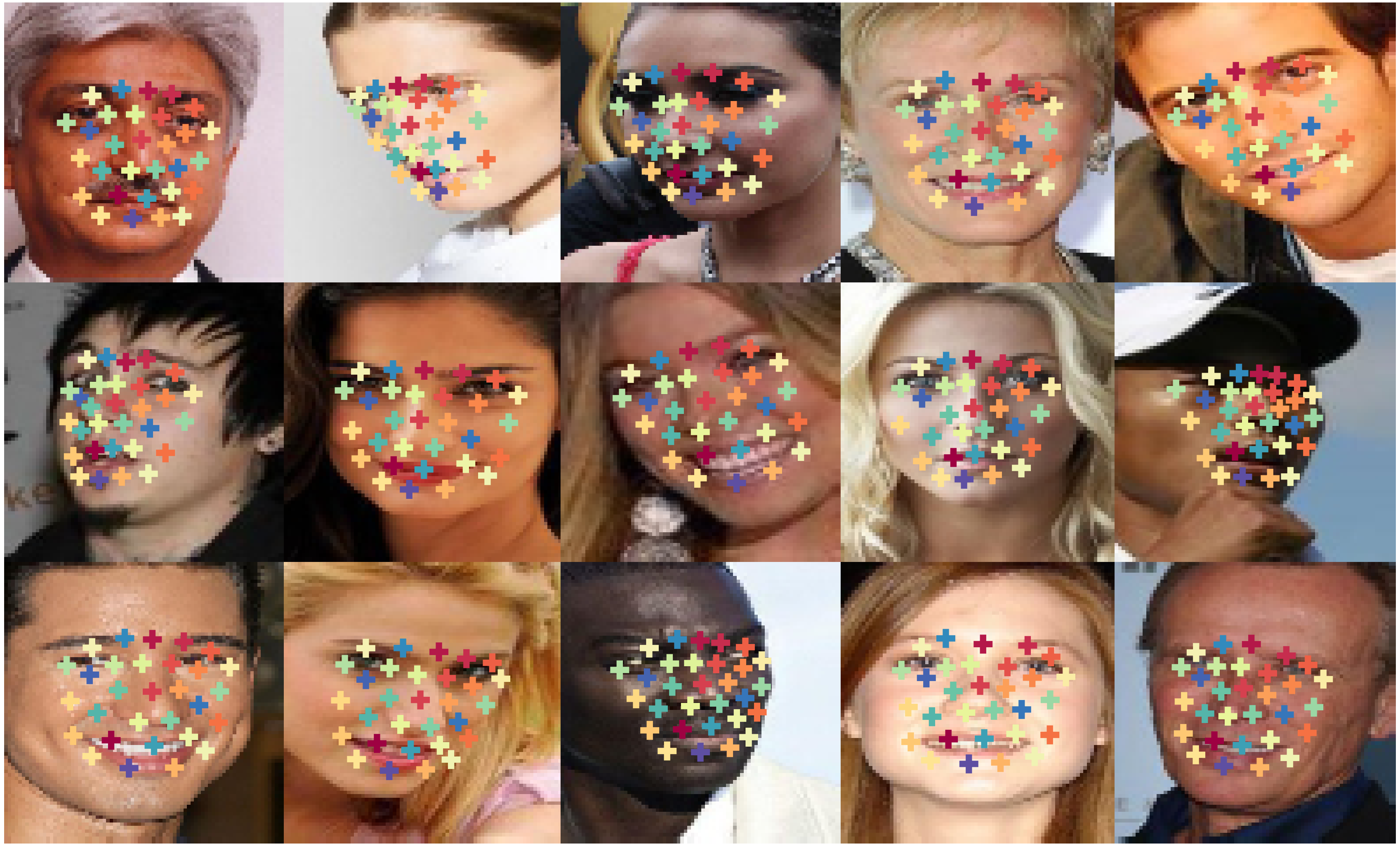


Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised discovery: Faces, 30 landmarks

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised discovery: Faces, 30 landmarks



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised landmark discovery: Cat head

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

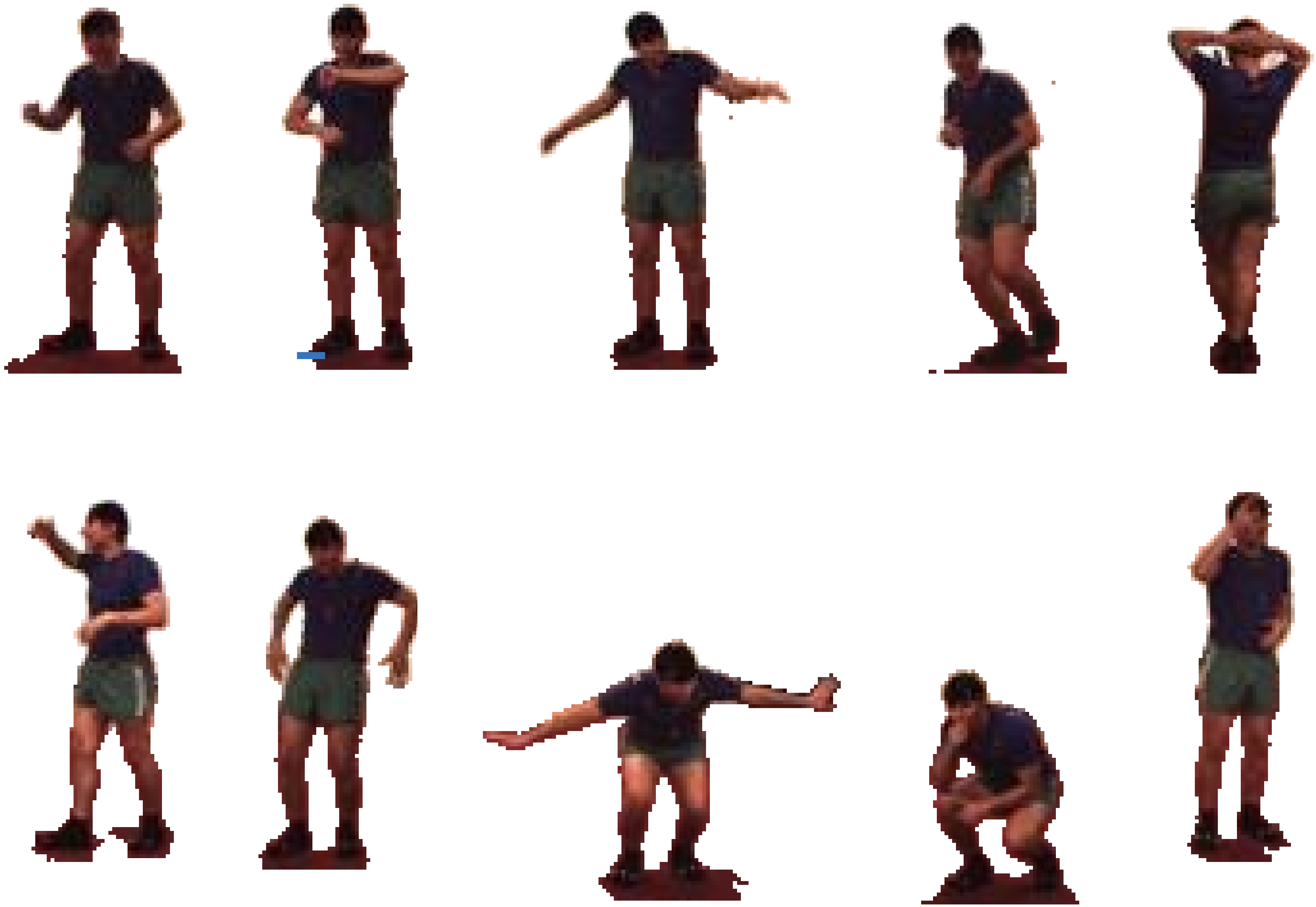
Unsupervised landmark discovery: Cat head

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised landmarks: shoes, cars, animals, MNIST

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised landmark discovery: Human3.6M

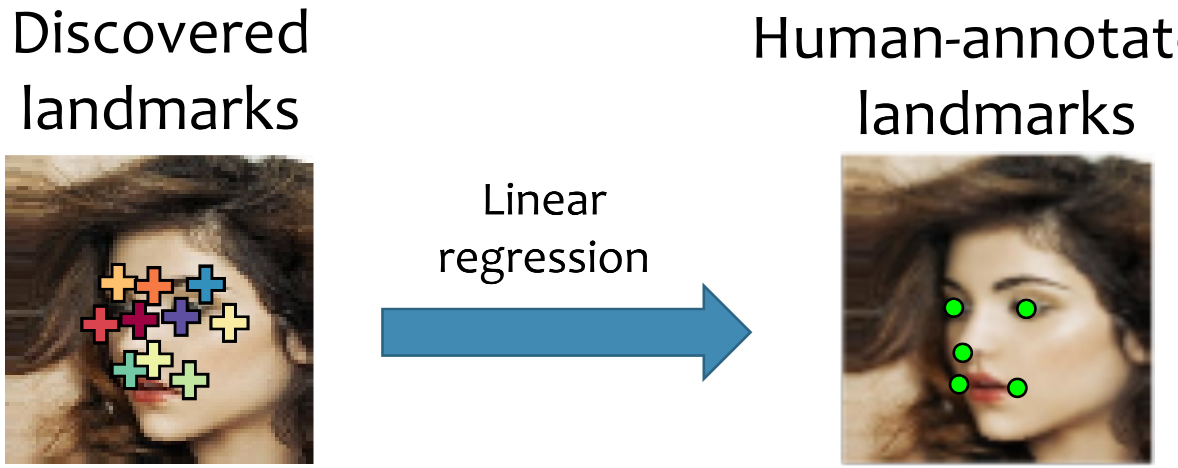


Unsupervised landmark discovery: Human3.6M

Optical flows for the equivariance

Quantitative evaluation: Regression to Ground Truth Landmarks

- Train a **linear regression model** to map the discovered landmark to human-annotated landmarks without finetuning the neural network.

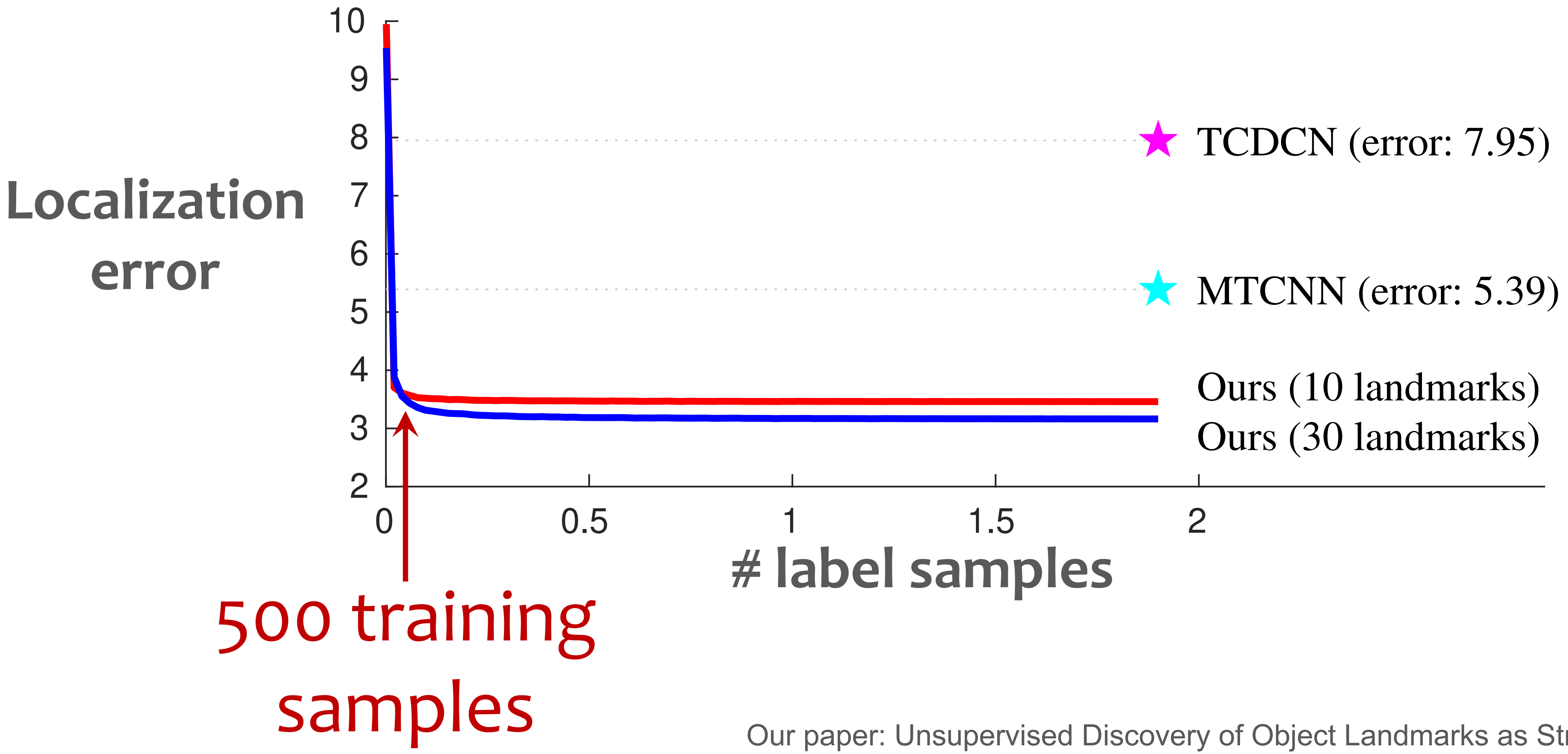


MAFL faces (5 target landmarks)

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

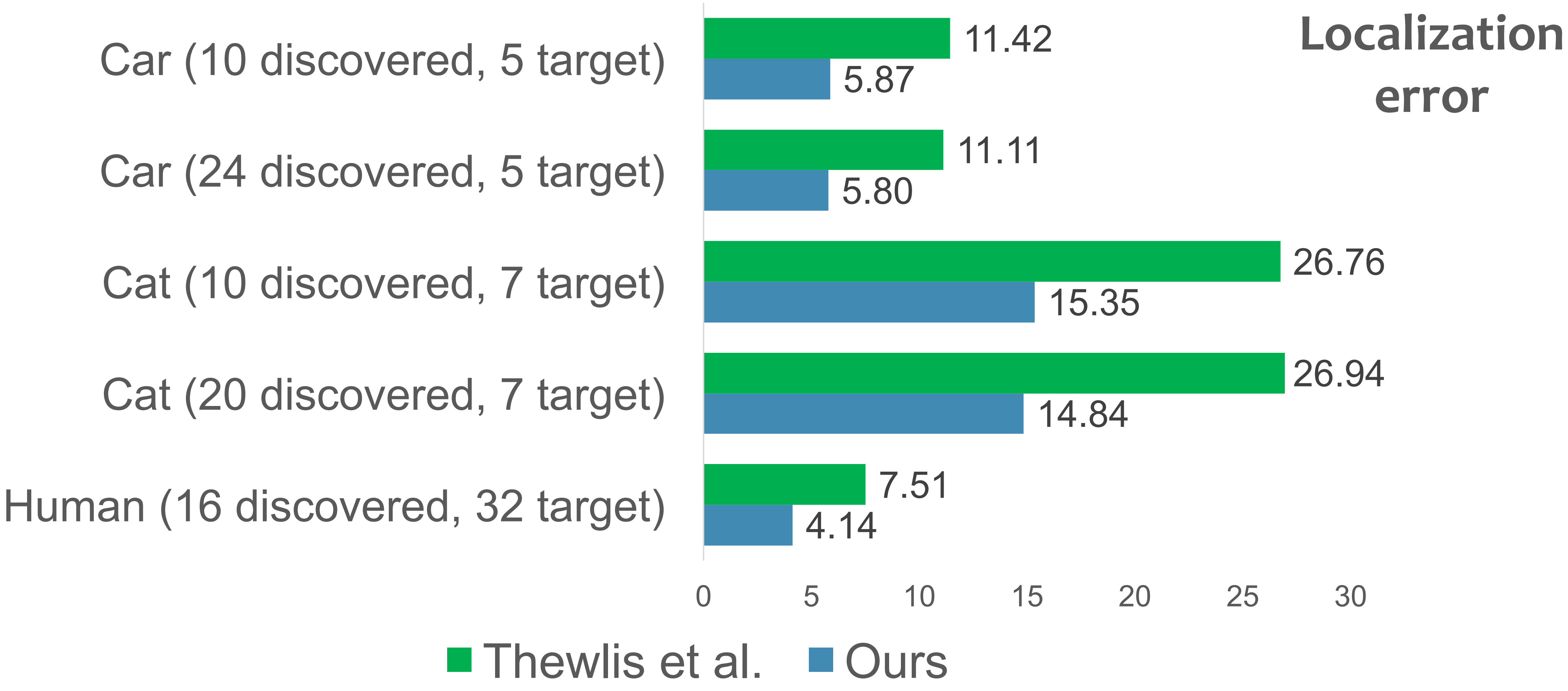
Semi-supervised learning

- Better landmark detector using less training samples



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Cars, cat heads, human bodies



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Facial attribute classification

- Landmark coordinates as visual representations
- Predicting 13 binary facial attributes that are related to the facial shape.

Arched Eyebrows, Bags Under Eyes, Big Lips, Big Nose, Double Chin, High Cheekbones, Male,
Mouth Slightly Open, Narrow Eyes, Oval Face, Pointy Nose, Receding Hairline, Smiling

Method	Feature dimension	Accuracy
Ours (discovered landmarks)	60	83.2

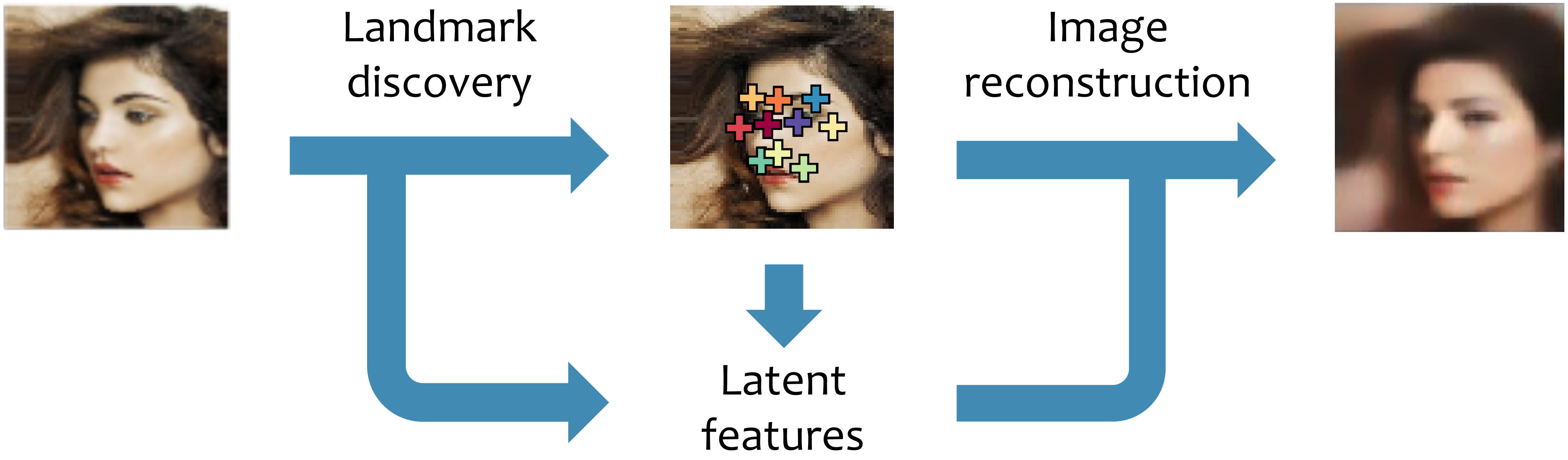
Facial attribute classification

- Landmark coordinates as visual representations
- Predicting 13 binary facial attributes that are related to the facial shape.

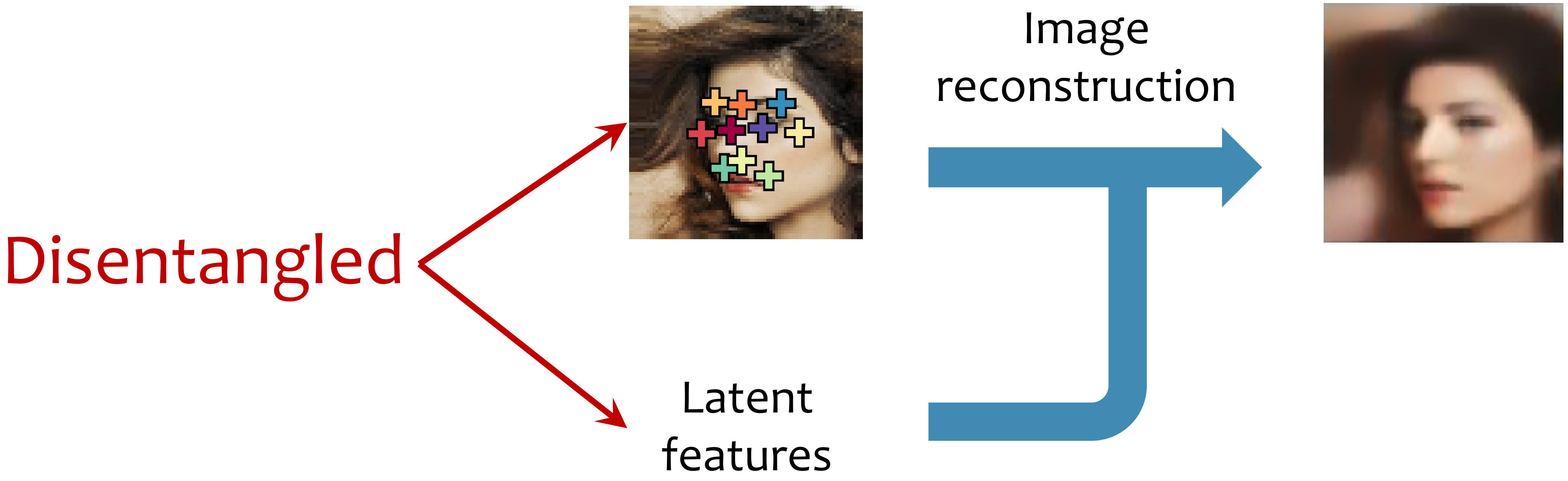
Arched Eyebrows, Bags Under Eyes, Big Lips, Big Nose, Double Chin, High Cheekbones, Male, Mouth Slightly Open, Narrow Eyes, Oval Face, Pointy Nose, Receding Hairline, Smiling

Method	Feature dimension	Accuracy
Ours (discovered landmarks)	60	83.2
FaceNet (top-layer)	128	80.0
FaceNet (conv-layer)	1792	82.4

[FaceNet] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015



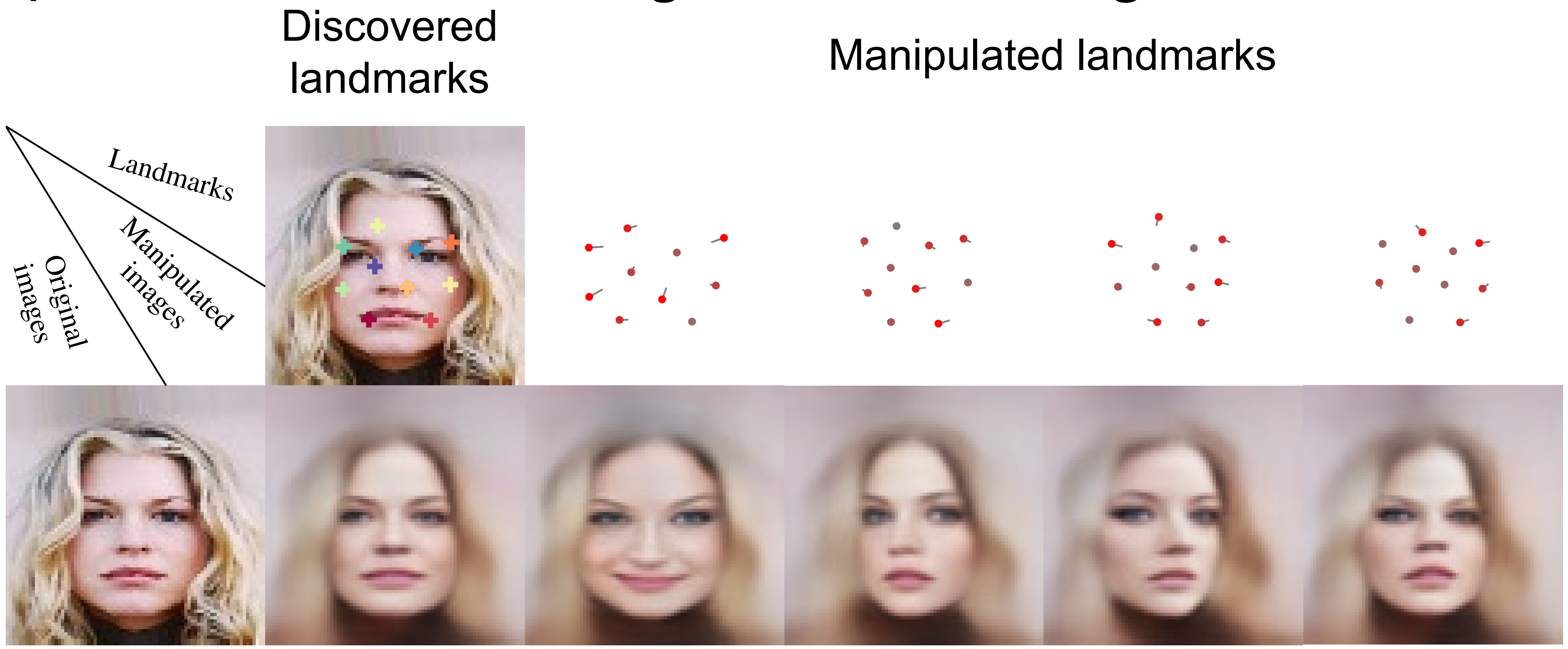
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Image manipulation

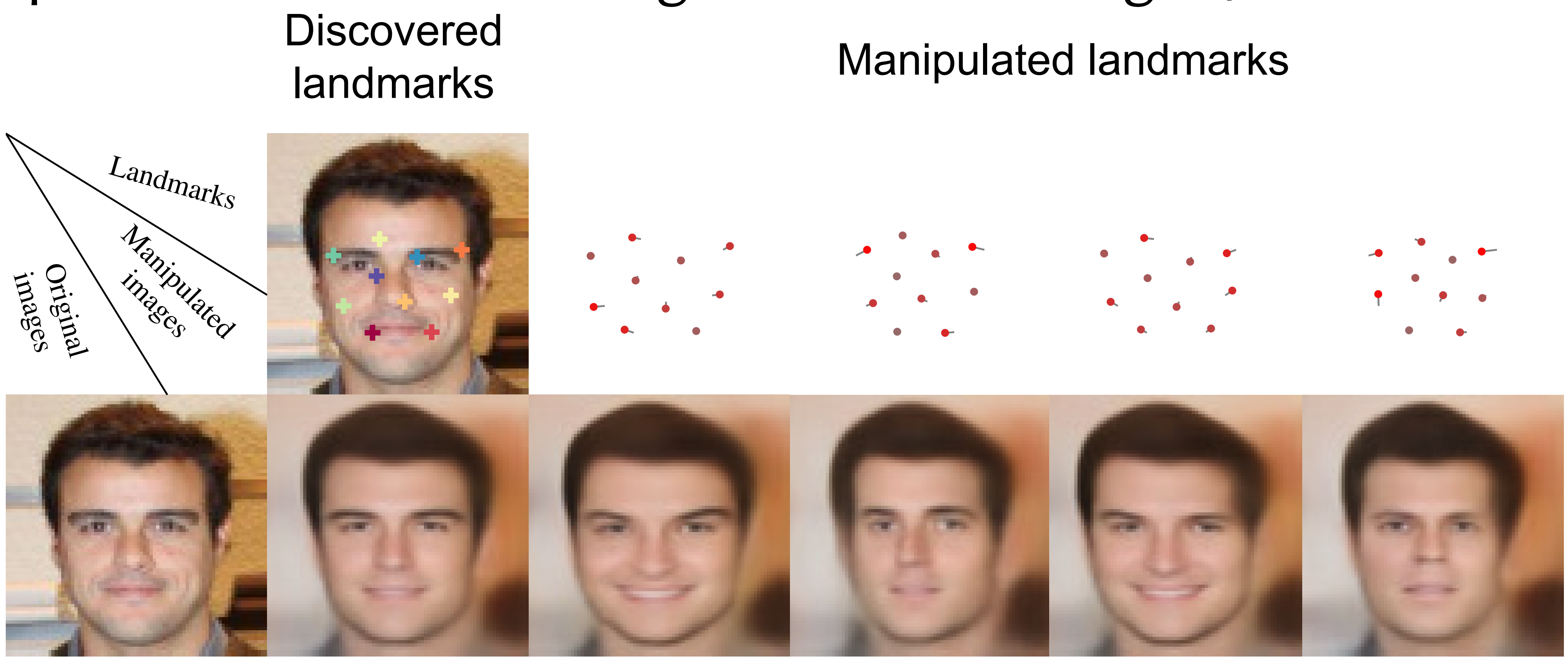
- Discover landmarks and extract latent features from an image.
- Manipulate the landmarks to generate new images / videos.



Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Image manipulation

- Discover landmarks and extract latent features from an image.
- Manipulate the landmarks to generate new images / videos.



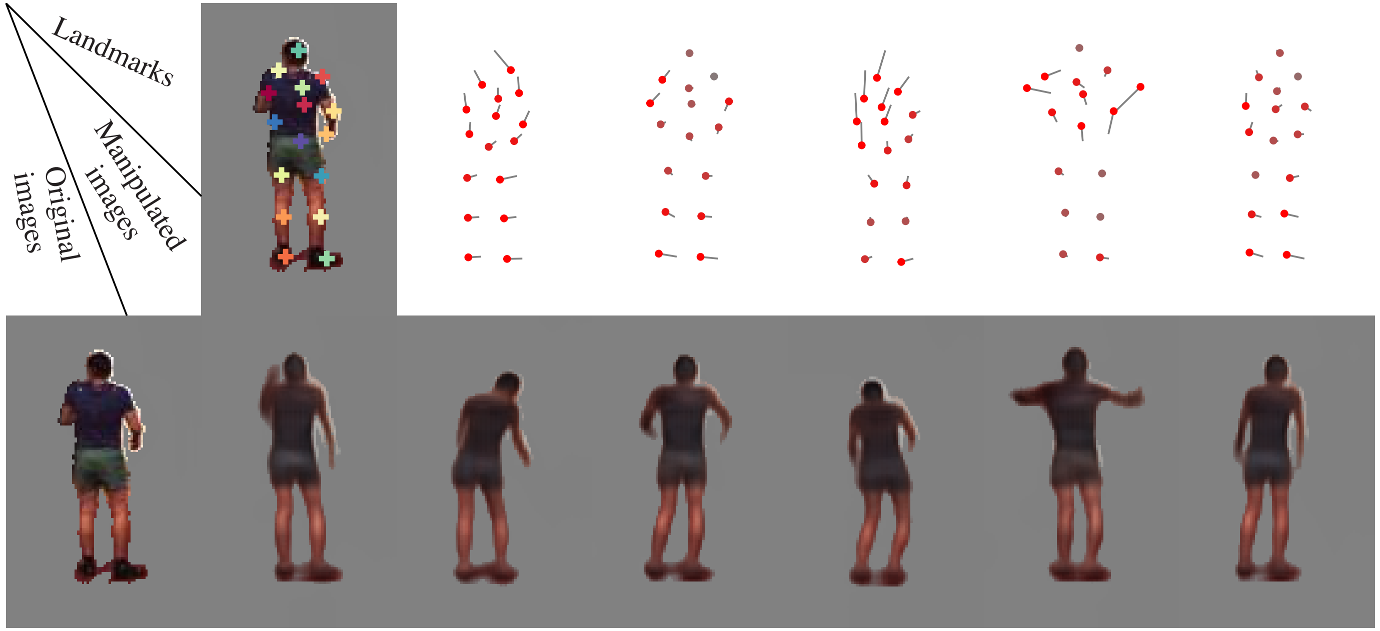
Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Image manipulation: Human body

- Discover landmarks and extract latent features from an image.
- Manipulate the landmarks to generate new images / videos.

Discovered landmarks

Manipulated landmarks



manipulating all 16 landmarks

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Conclusions

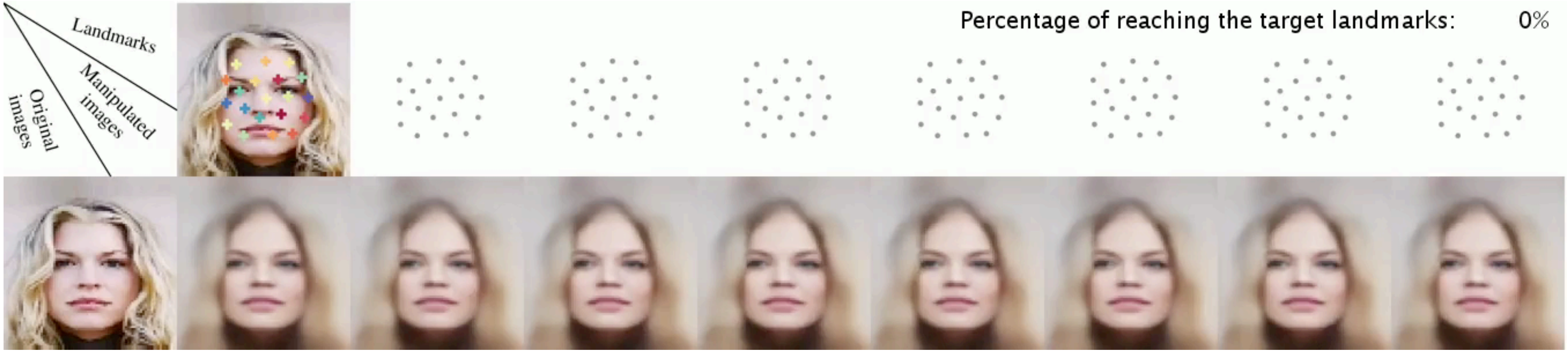
- Unsupervised object landmark discovery as image representations with explicit structures
- A fully differentiable neural network architecture

Conclusions

- Unsupervised object landmark discovery as image representations with explicit structures
- A fully differentiable neural network architecture
- Our unsupervised model can
  - produce meaningful landmarks
  - perform competitively to supervised facial landmark detector
  - provide a neural-network interface that humans can manipulate

Our paper: Unsupervised Discovery of Object Landmarks as Structural Representations

Unsupervised Discovery of Object Landmarks as Structural Representations



Poster E19

Thank you!

Project page
(Code & results):

http://ytzhang.net/projects/lmdis-rep