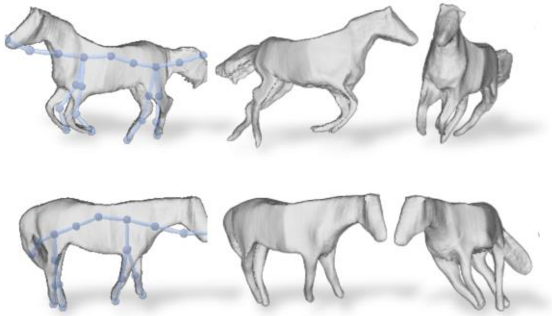
Unsupervised 3D Structure Inference from Category-Specific Image Collections

Weikang Wang   Dongliang Cao   Florian Bernard
University of Bonn

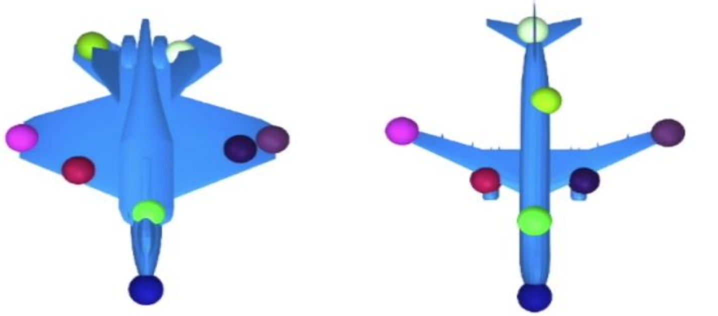
Background & Motivation

- 3D structure represented by keypoints (possibly with linking edges) are useful for many downstream tasks, such as:

Graph matching (from [1])



3D Shape animation (from [2])



Control 3D generation (from [3])

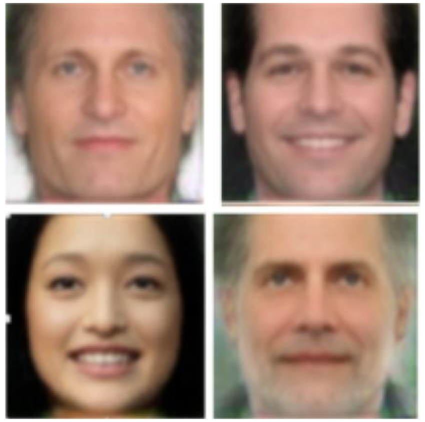
[1] *Nurlanov, Z., Schmidt, F. R., & Bernard, F. Universe points representation learning for partial multi-graph matching. AAAI 2023*
[2] *Wu, S., Li, R., Jakab, T., Rupprecht, C., & Vedaldi, A. Magicpony: Learning articulated 3d animals in the wild. CVPR 2023*
[3] *Jakab, T., Tucker, R., Makadia, A., Wu, J., Snavely, N., & Kanazawa, A. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. CVPR 2023*

Background & Motivation

- However, inferring 3D keypoints from images is hard, and previous unsupervised works use various priors to achieve this:

Multiple views [1, 2, 3]

2D keypoints annotations (various SfM methods [4,5], [6])



Geometry constraints (such as symmetry [7], skeleton [8])

[1] *Chen, B., Abbeel, P., & Pathak, D. Unsupervised learning of visual 3d keypoints for control. ICML 2021*

[2] *Honari, S., & Fua, P. Unsupervised 3d keypoint estimation with multi-view geometry. Arxiv 2022*

[3] *Suwajanakorn, S., Snavely, N., Tompson, J. J., & Norouzi, M. (2018). Discovery of latent 3d keypoints via end-to-end geometric reasoning. NeurIPS 2018*

[4] *Kong, C., & Lucey, S. Deep non-rigid structure from motion. CVPR 2019*

[5] *Novotny, D., Ravi, N., Graham, B., Neverova, N., & Vedaldi, A. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. ICCV 2019*

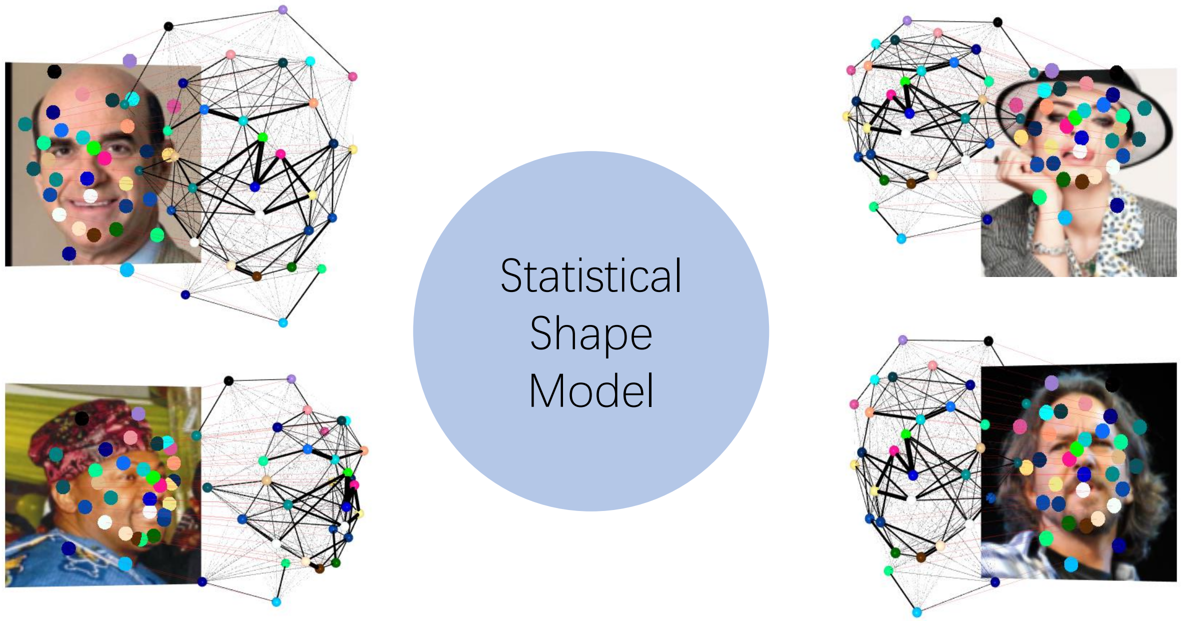
[6] *Reddy, N. D., Vo, M., & Narasimhan, S. G. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. CVPR 2019*

[7] *Wu, S., Rupprecht, C., & Vedaldi, A. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. CVPR 2020*

[8] *He, X., Bharaj, G., Ferman, D., Rhodin, H., & Garrido, P. Few-shot geometry-aware keypoint localization. CVPR 2023*

Background & Motivation

- In this paper, we inference 3D keypoints directly from a **category-specific** image collection (no multiple views) **without** any priors.

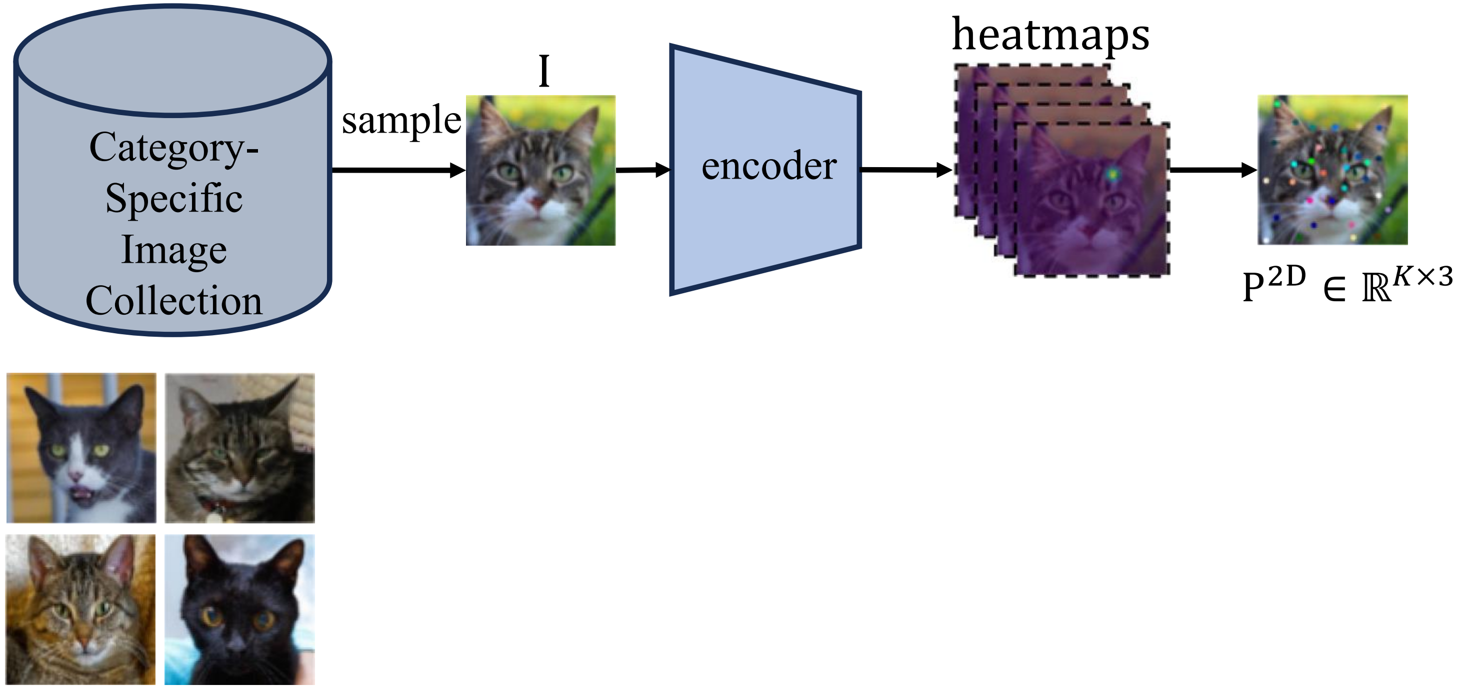


- The core idea is: *Different instances from a same category share a* ***similar*** *sparse 3D structure with* ***restricted*** *deformations.*

Method

- Inputs: a set of category-specific image collections

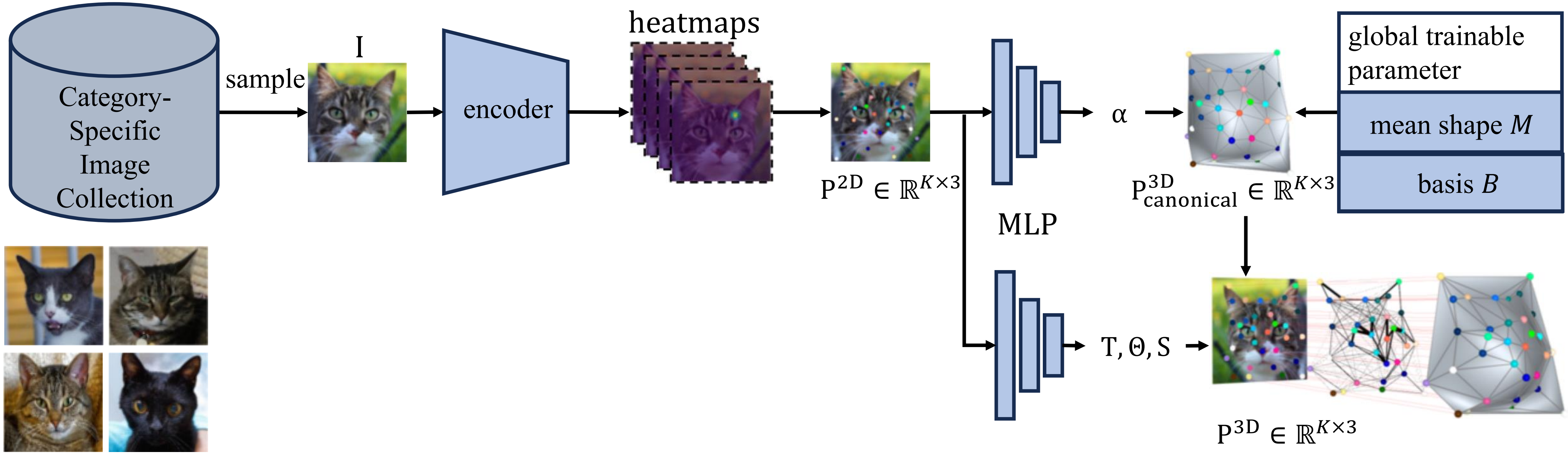
Method



Sample an image I and feed it into the encoder to get K heatmaps $H_i \in [0,1]^{\mathrm{H}\times\mathrm{W}}$, $i = 1, \cdots K$, then get the 2D keypoint matrix $\mathrm{P}^{2\mathrm{D}} \in \mathbb{R}^{K\times 3}$ via:

$$\mathrm{P}_i^{2\mathrm{D}} = \sum_p \frac{H_i(p)}{\sum_{p'} H_i(p')} p, \qquad i = 1, \cdots, K, \qquad p, p' \in [-1,1] \times [-1,1]$$

Method

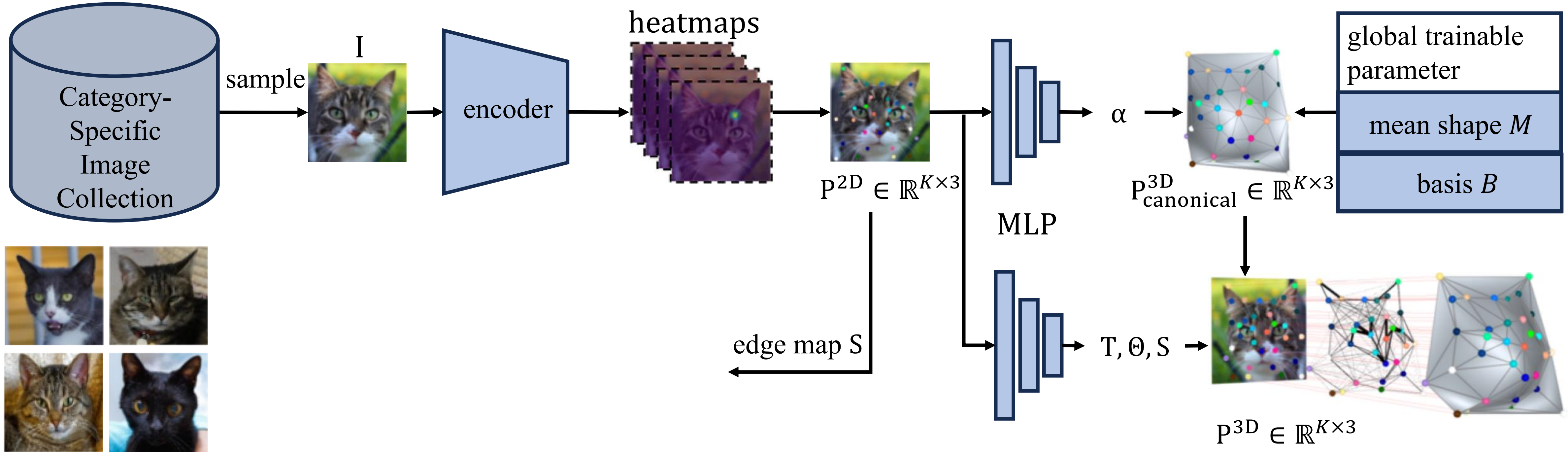


Meanwhile, global trainable parameters mean shape $M \in \mathbb{R}^{K\times3}$, and a set of basis $B \in \mathbb{R}^{n\times3K}$ are learned. Together with basis coefficient α, we get the 3D keypoints at canonical pose:

$$P^{3D}_{canonical} = M + \alpha B$$

Then P^{3D} is got by apply rigid body transformation (R, T) and scaling factor S on $P^{3D}_{canonical}$

Method



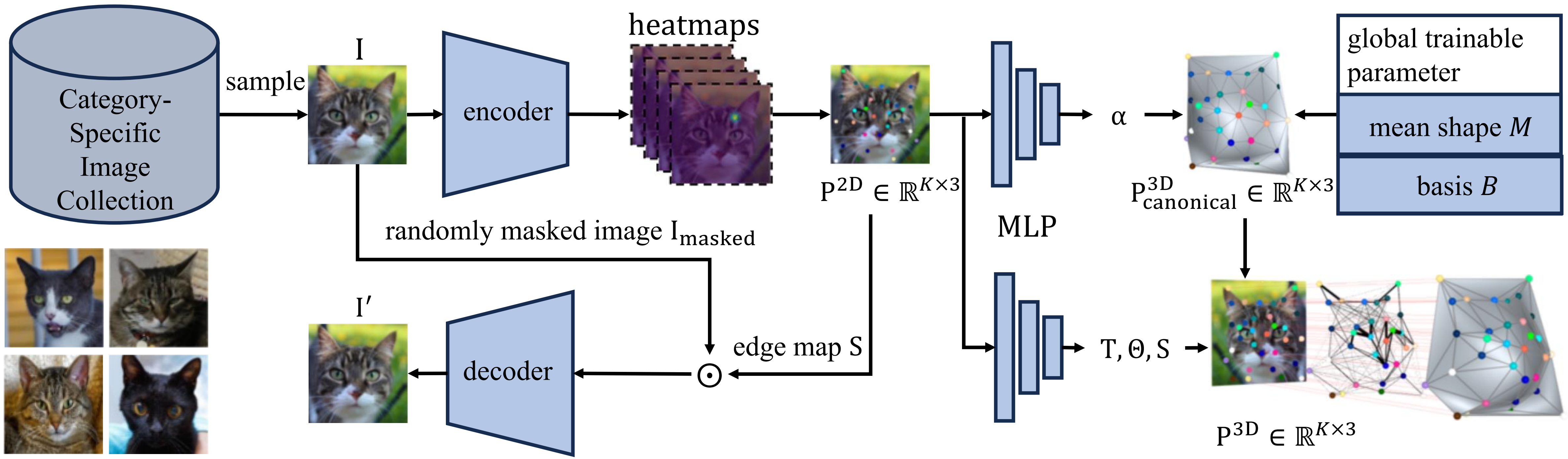
For each pair of 2D keypoints, P_i^{2D} and P_j^{2D}, an edge map $S_{ij} \in \mathbb{R}^{H\times W}$ is defined on normalized pixel coordinates p:

$$S_{ij} = exp\left(-\frac{d_{ij}(p)}{\sigma}\right), \qquad 1 \le i,j \le K$$

where $d_{ij}(p)$ is the distance of pixel p to line segment connecting P_i^{2D} and P_j^{2D}, and $\sigma \in \mathbb{R}$ controls the thickness of the edge. The final edge map $S \in \mathbb{R}^{H\times W}$ summarizing S_{ij} for paired keypoints is:

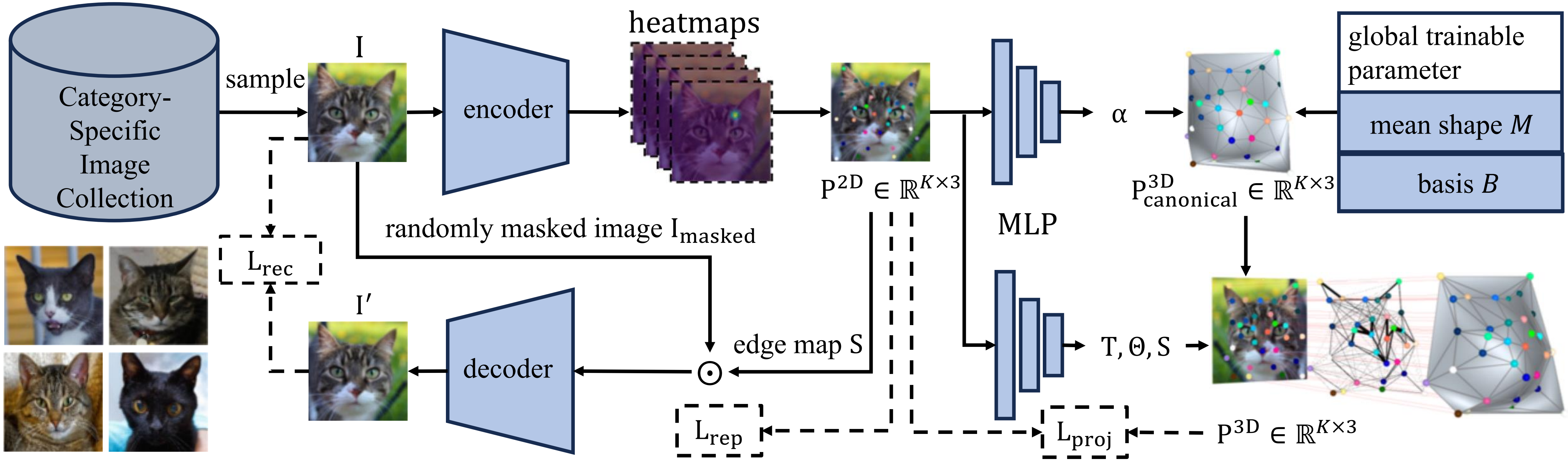
$$S(p) = \max_{1\le i,j\le K} \omega_{ij} S_{ij}(p)$$

Method



sample
Category-Specific Image Collection
I
encoder
heatmaps
$P^{2D} \in \mathbb{R}^{K\times3}$
MLP
α
$P^{3D}_{canonical} \in \mathbb{R}^{K\times3}$
global trainable parameter
mean shape M
basis B
randomly masked image I_{masked}
I′
decoder
⊙
edge map S
T, Θ, S
$P^{3D} \in \mathbb{R}^{K\times3}$

Method



- **Reconstruction loss:** $\mathcal{L}_{rec} = \left\| F\left(D(S \odot I_{masked})\right) - F(I) \right\|$
- **Projection loss:** $\mathcal{L}_{proj} = \left\| P^{3D}\Pi - P^{2D} \right\|$, *where* $\Pi = [s_1, 0\,;\ 0, s_2\,; 0, 0] \in \mathbb{R}_{+}^{3\times 2}$
- **Repulsion loss:** $\mathcal{L}_{rep} = -\sum_{i=1}^{K} \left\| P_i^{2D} - \mathcal{N}_i \right\| \exp\left(-\frac{\left\| P_i^{2D} - \mathcal{N}_i \right\|}{h}\right)$, *where* $\mathcal{N}_i$ is nearest to P_i^{2D}

Quantitative Results

Method	K=8
DFF [1]	31.30%*
SCOPS (w/o saliency) [6]	22.11%†
SCOPS (w/saliency) [6]	15.01%†
Liu et al. [8]	12.26%†
Huang et al. [5]	8.4%†
GANSeg [4]	6.18%†
Thewlis et al. [10]	31.30%*
Zhang et al. [11]	40.82%*
LatentKeypointGAN [2]	21.90%†
LatentKeypointGAN-tuned [2]	5.63%†
Lorenz et al. [9]	11.41%‡
IMM [7]	8.74%‡
AutoLink [3]	5.39%
Ours	**5.21%**

Tab1. Normalized L_2 error (NME) for 2D keypoints inference of various unsupervised methods on CELEBA WILD datasets for $K = 8$

Method	K=8	K=16	K=24	K=32
AutoLink [3]	5.39%	4.69%	3.99%	3.77%
Ours	**5.21%**	**3.97%**	**3.54%**	**3.48%**

Tab2. Normalized L_2 error (NME) of AutoLink and our method for different numbers of keypoints using the CELEBA WILD dataset.

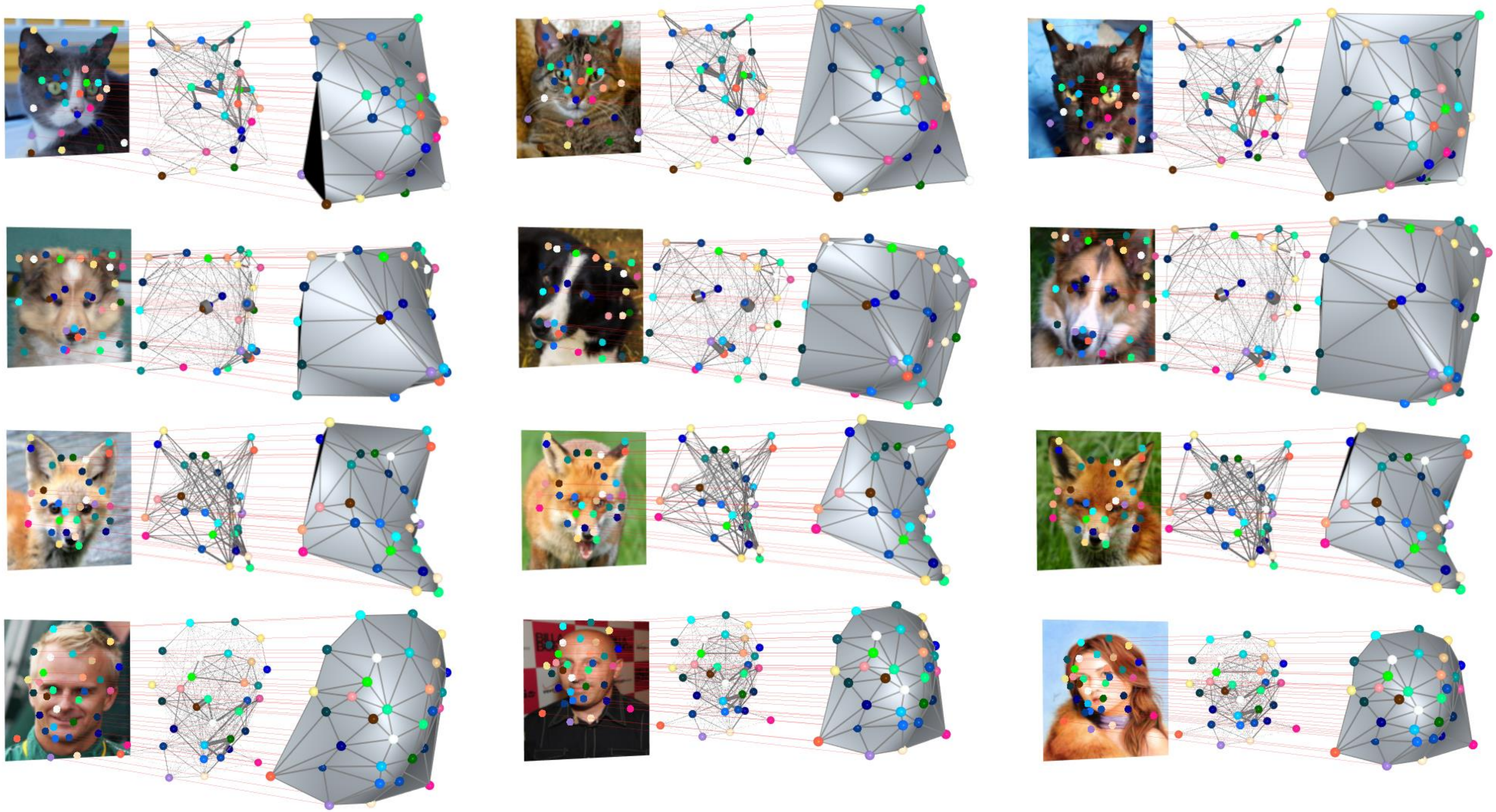
Supervised	Unsupervised		
3DDFA	AutoLink+Unsup3d	AutoLink+MiDaS	Ours
4.94%	11.47%	9.23%	**8.48%**

Tab3. Normalized L_2 error (NME) of our method, two unsupervised methods (AutoLink + MiDaS and AutoLink + Unsup3d) and one supervised method (3DDFA) for 3D keypoints inference (training on the 300W-LP dataset and testing on the AFLW2000- 3D dataset).

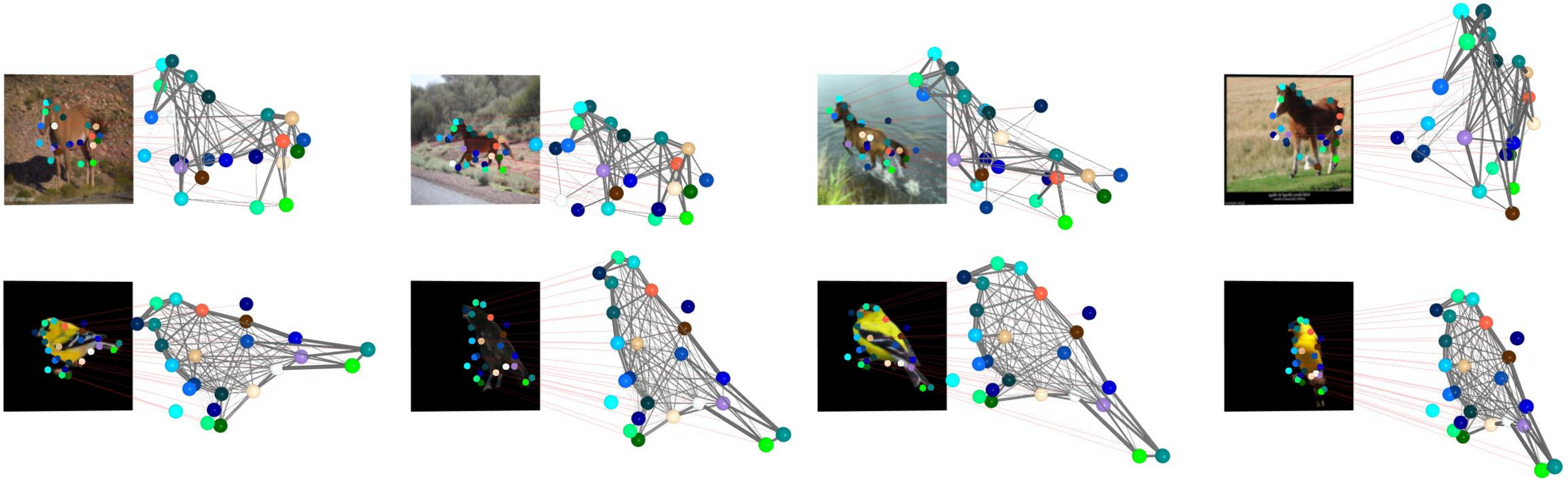
References

[1] Collins, E., Achanta, R., & Susstrunk, S., Deep feature factorization for concept discovery. ECCV 2018
[2] He, X., Wandt, B., & Rhodin, H., Latentkeypointgan: Controlling GANs via latent keypoints. Arxiv 2021
[3] He, X., Wandt, B., & Rhodin, H., Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. NeurIPS 2022
[4] He, X., Wandt, B., & Rhodin, H., Ganseg: Learning to segment by unsupervised hierarchical image generation. CVPR 2022
[5] Huang, Z., & Li, Y., Interpretable and accurate fine grained recognition via region grouping. CVPR2020
[6] Hung, W. C., Jampani, V., Liu, S., Molchanov, P., Yang, M. H., & Kautz, J., Scops: Self-supervised co-part segmentation. CVPR 2019
[7] Jakab, T., Gupta, A., Bilen, H., & Vedaldi, A., Unsupervised learning of object landmarks through conditional image generation. NeurIPS 2018
[8] Liu, S., Zhang, L., Yang, X., Su, H., & Zhu, J., Unsupervised part segmentation through disentangling appearance and shape. CVPR 2021
[9] Lorenz, D., Bereska, L., Milbich, T., & Ommer, B., Unsupervised part-based disentangling of object shape and appearance. CVPR 2019
[10] Thewlis, J., Bilen, H., & Vedaldi, A., Unsupervised learning of object landmarks by factorized spatial embeddings. ICCV 2017
[11] Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., & Lee, H., Unsupervised discovery of object landmarks as structural representations. CVPR 2018
[12] Zhu, X., Liu, X., Lei, Z., & Li, S. Z., Face Alignment in Full Pose Range: A 3D Total Solution. TPAMI 2017

Qualitative Results



Qualitative Results



Thank you for listening!