Unsupervised Learning of Keypoints through Conditional Image Generation

Tomas Jakab* [1], Ankush Gupta* [1], Hakan Bilen[2], Andrea Vedaldi[1]

[1] VGG, University of Oxford, [2] University of Edinburgh

* equal contribution

arxiv.org/abs/1806.07823

Our goal

Learn semantically meaningful landmarks without any manual annotations

Motivation

Why to learn landmarks?

Low dimensional object representation

Interpretable

Why unsupervised?

Reduce dependency on expensive manual annotations

Leverage vast amount of videos available online

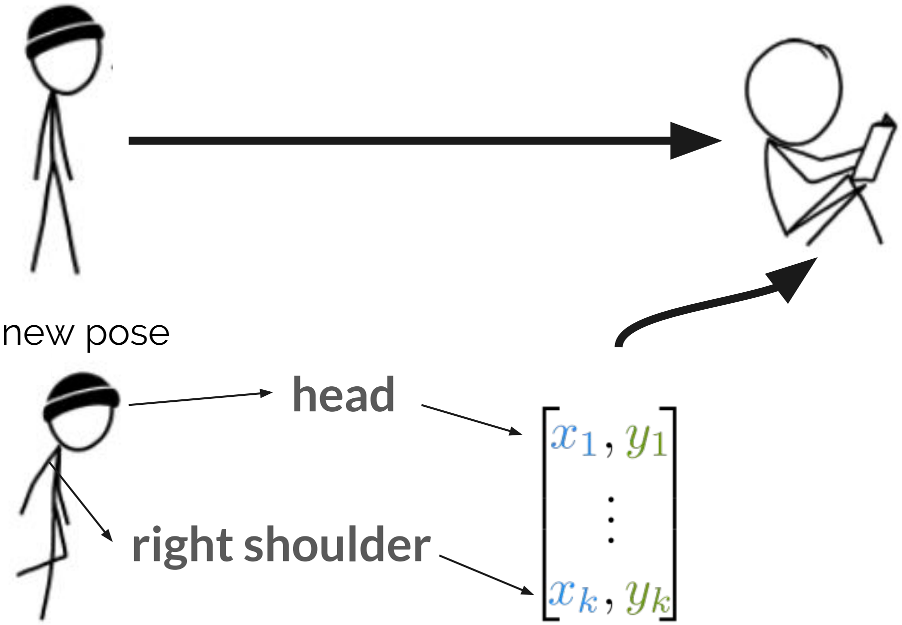
Method

Intuition

Intuition

new pose

Intuition



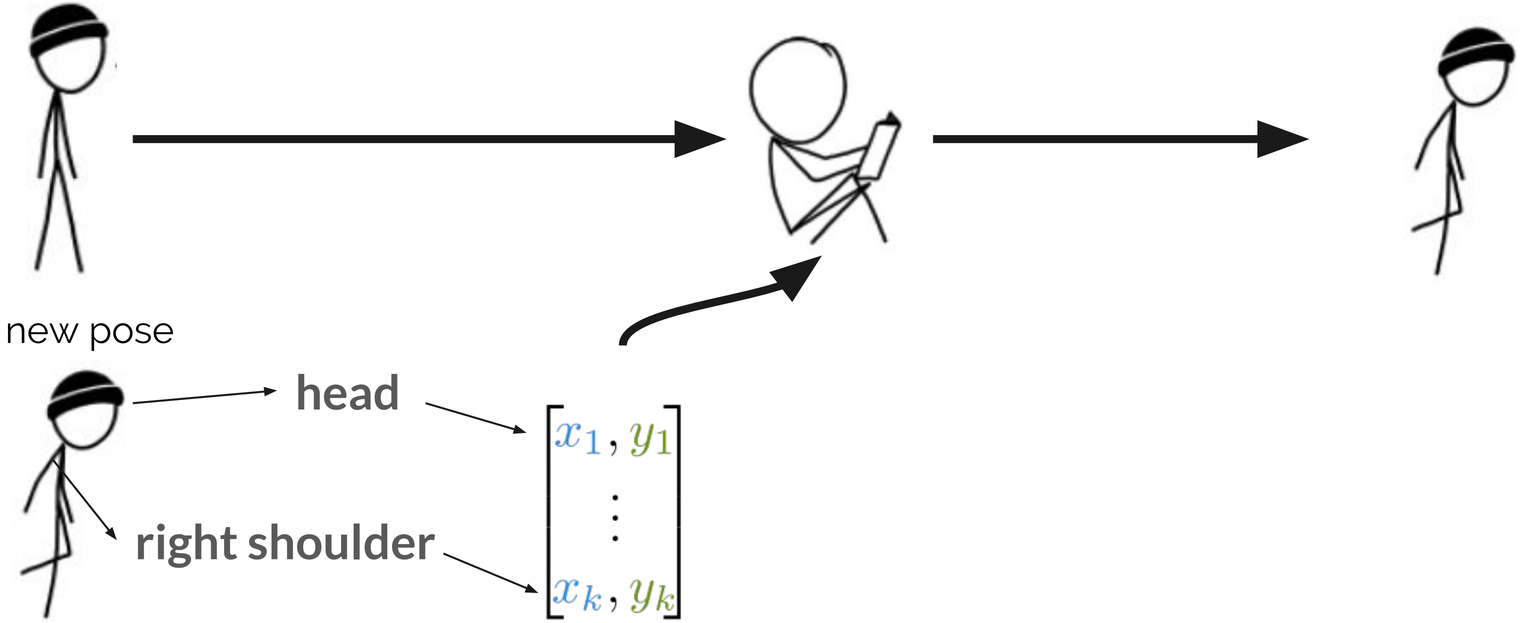
new pose

head

right shoulder

$$\begin{bmatrix} x_1, y_1 \\ \vdots \\ x_k, y_k \end{bmatrix}$$

Intuition



new pose

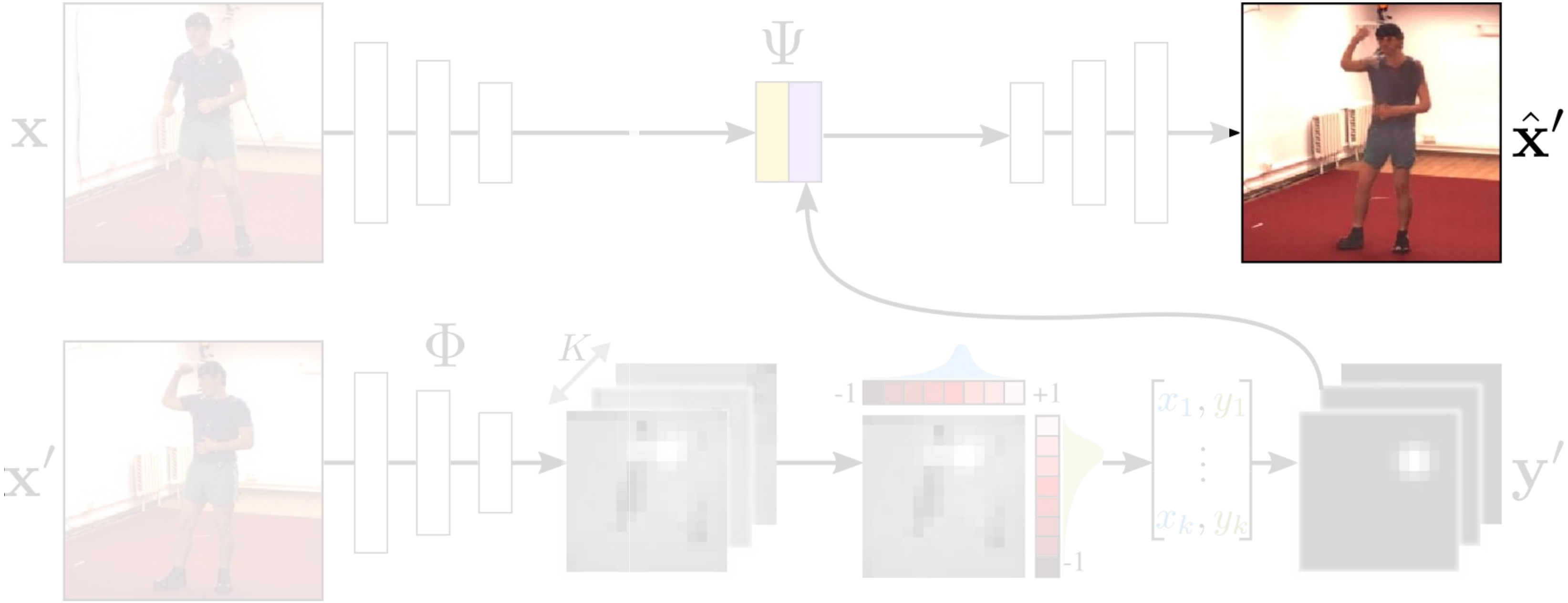
head

right shoulder

$$\begin{bmatrix} x_1, y_1 \\ \vdots \\ x_k, y_k \end{bmatrix}$$

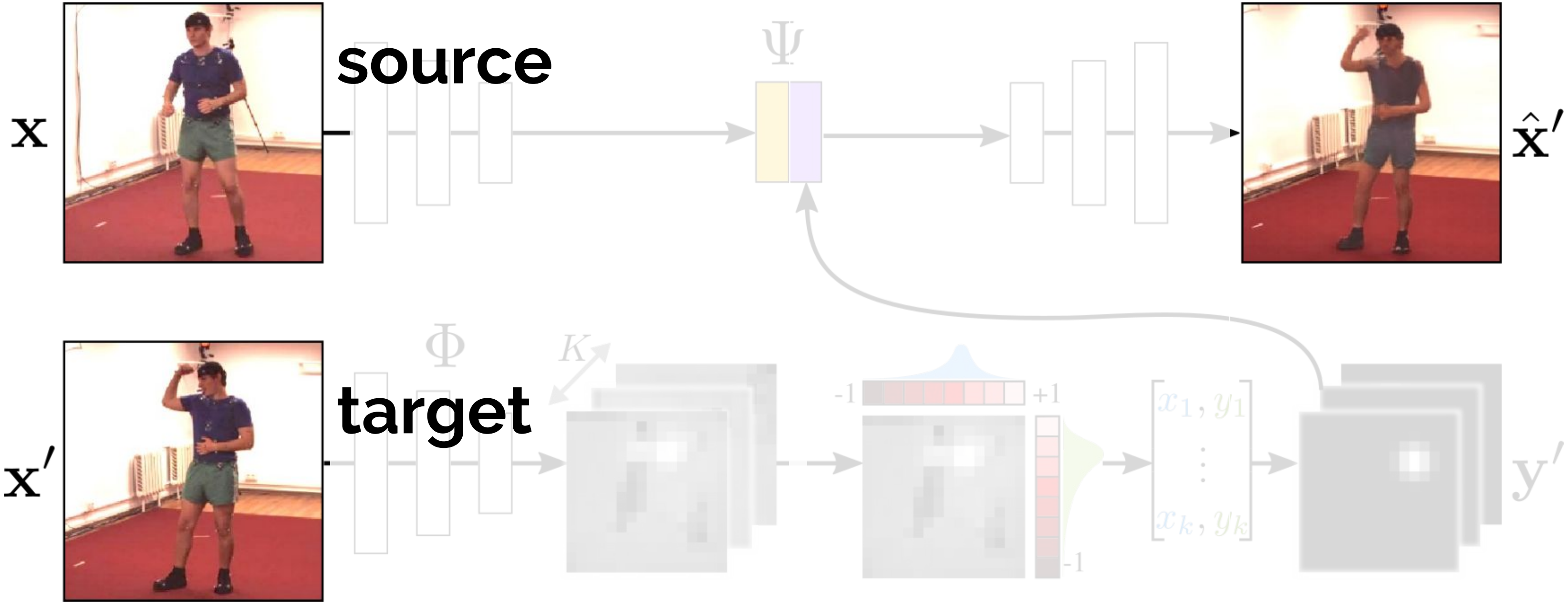
Model

goal: reconstruct target



Model

goal: reconstruct target

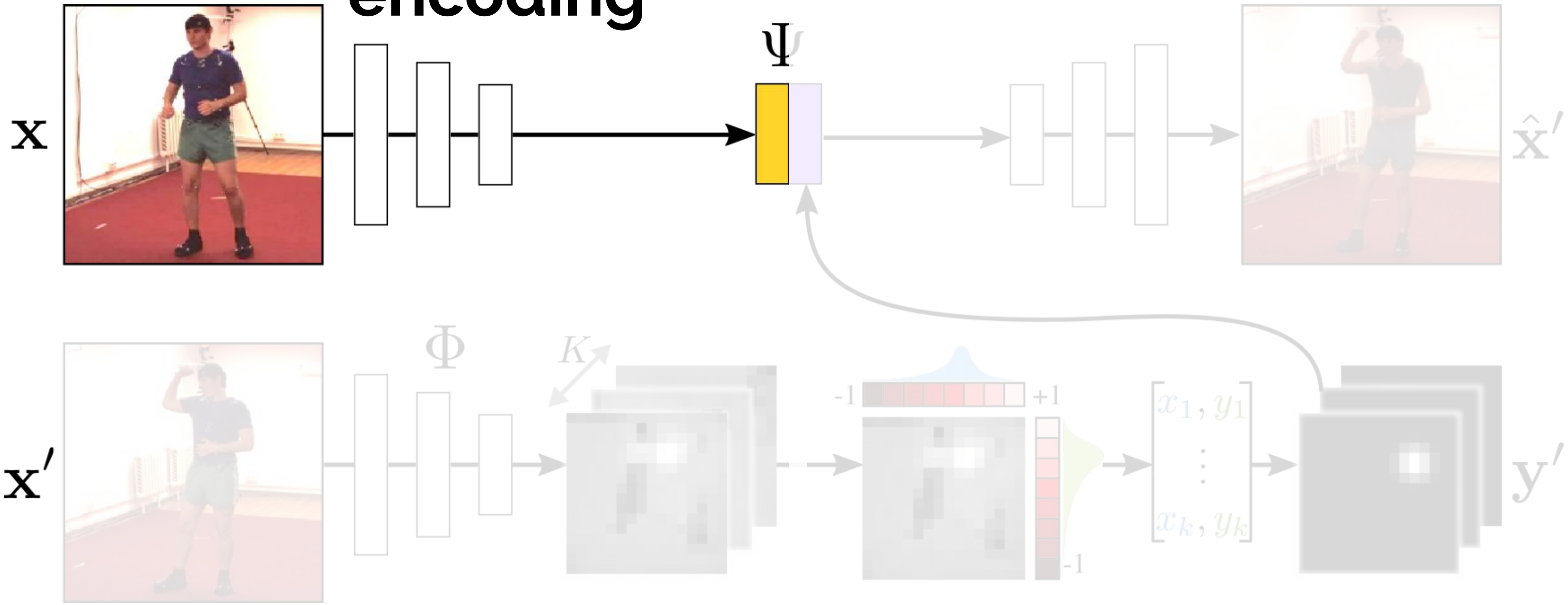


source

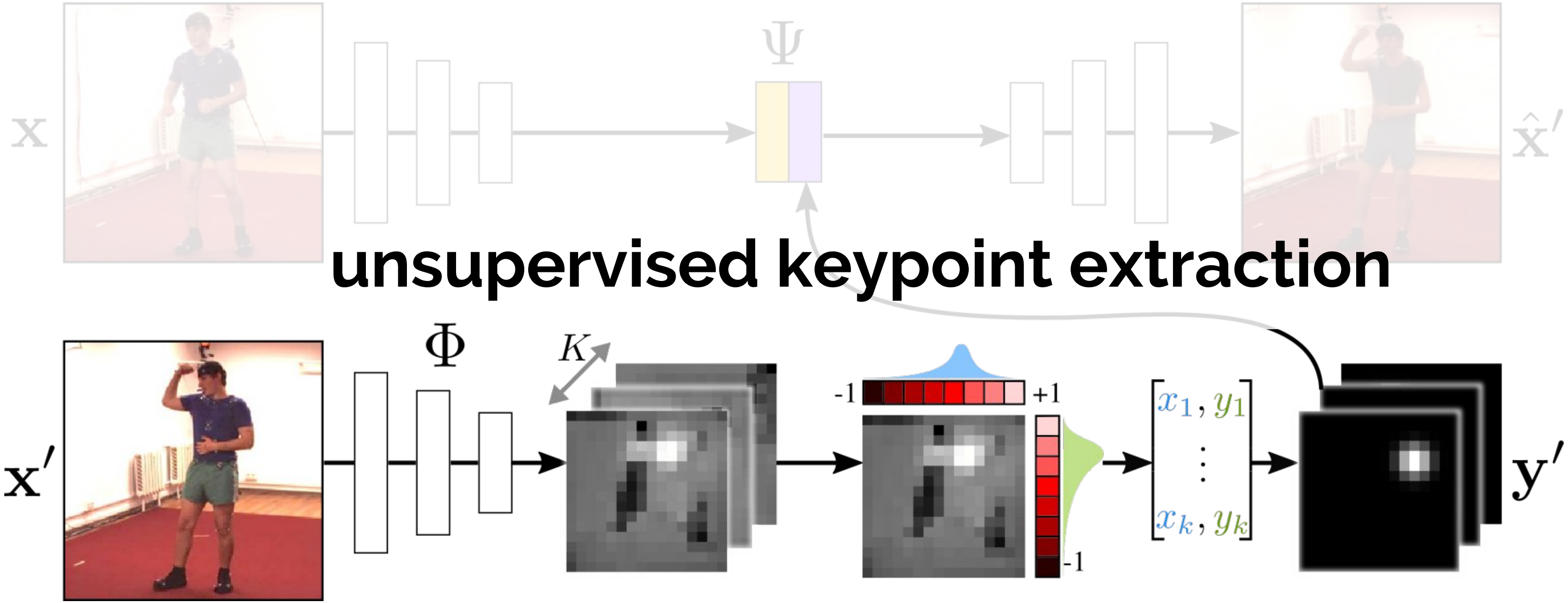
target

Model

appearance encoding



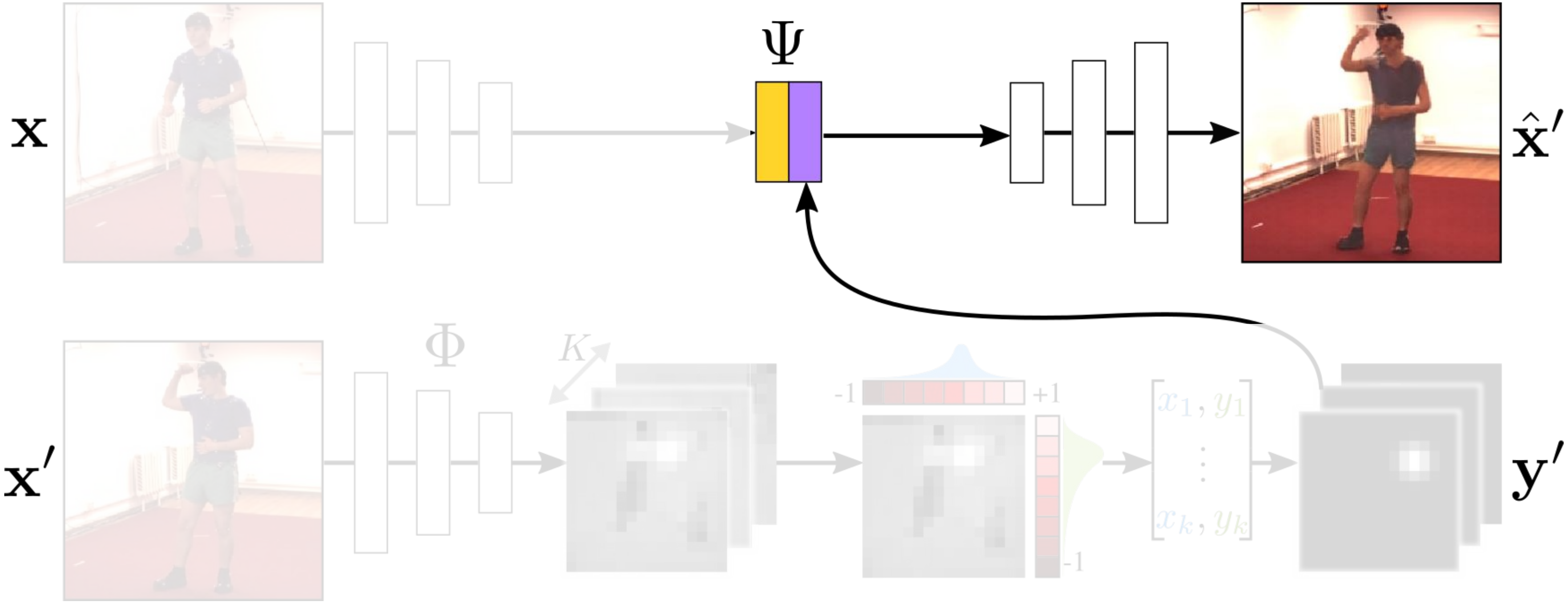
Model



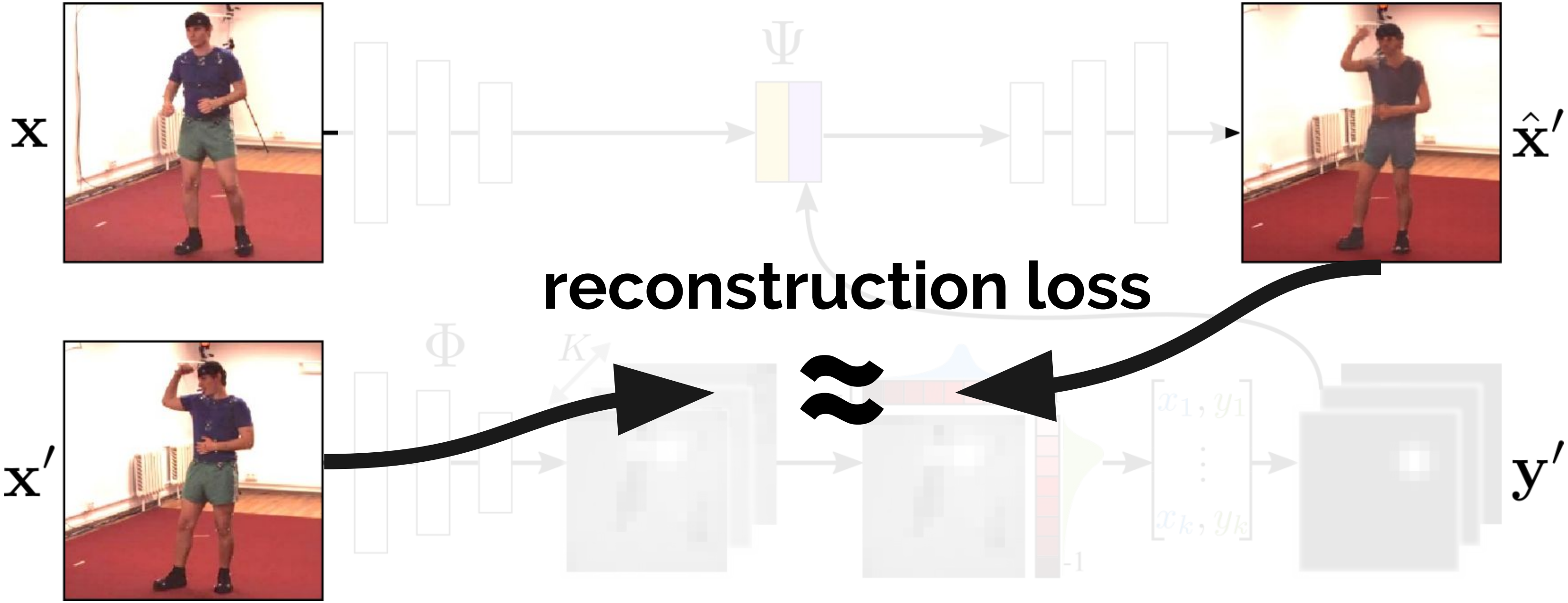
unsupervised keypoint extraction

Model

image reconstruction

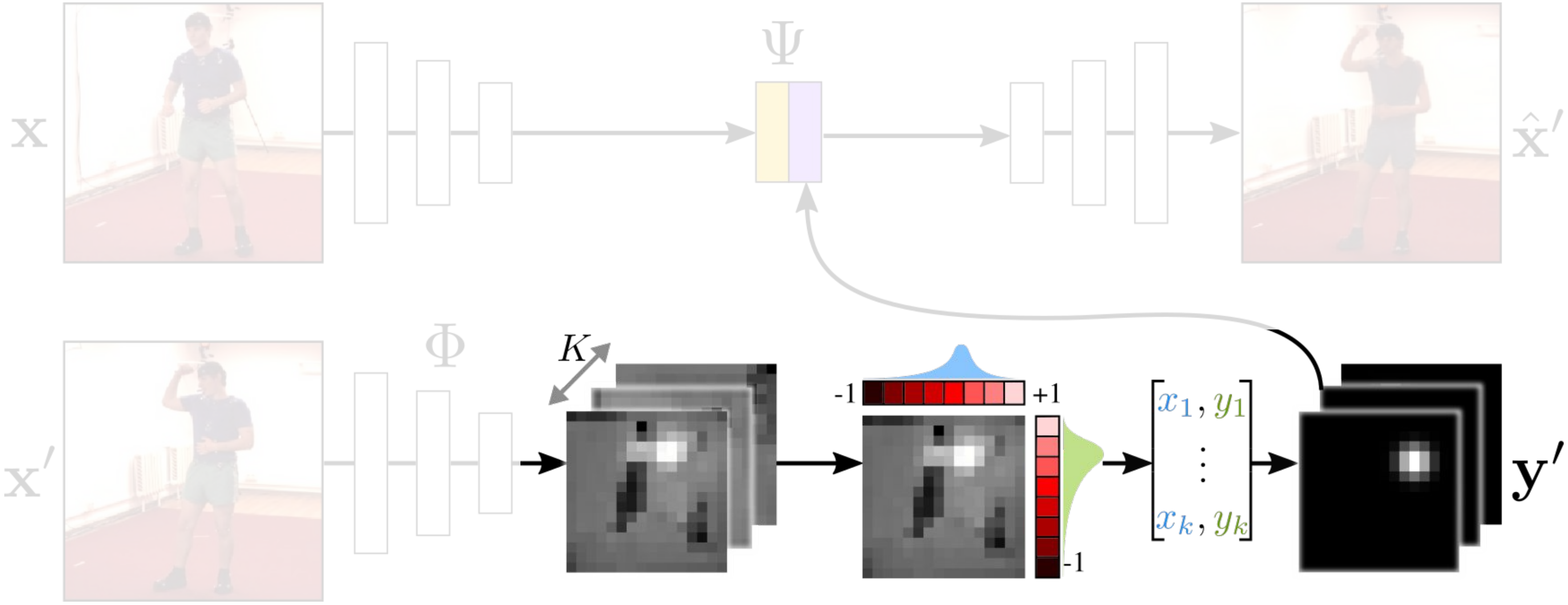


Loss

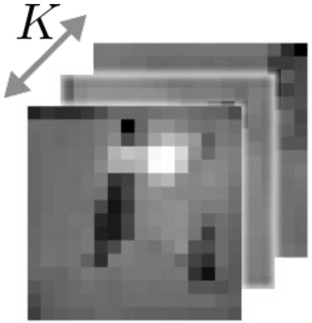


reconstruction loss

Bottleneck



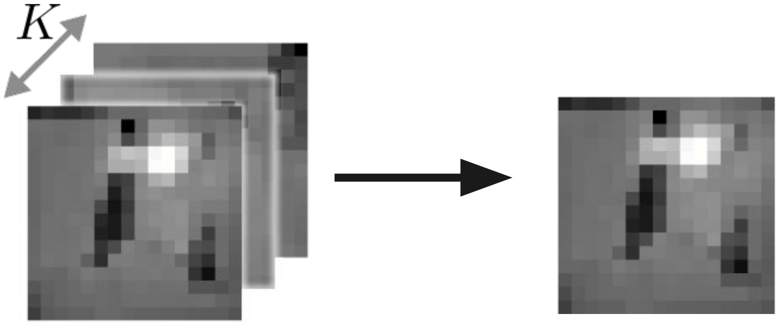
Bottleneck



heatmap for each keypoint

Bottleneck

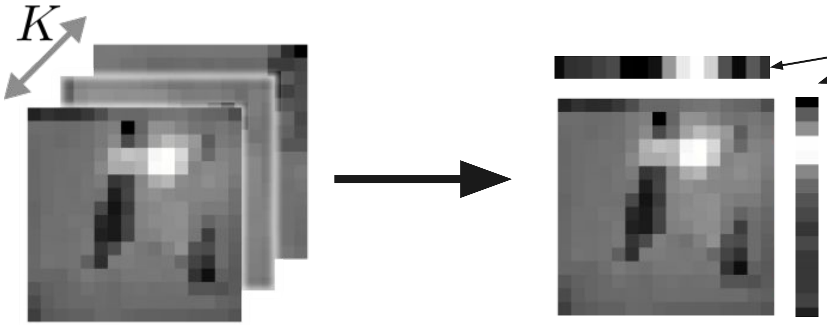
for each heatmap



heatmap for each keypoint

row-wise & column-wise sum

Bottleneck



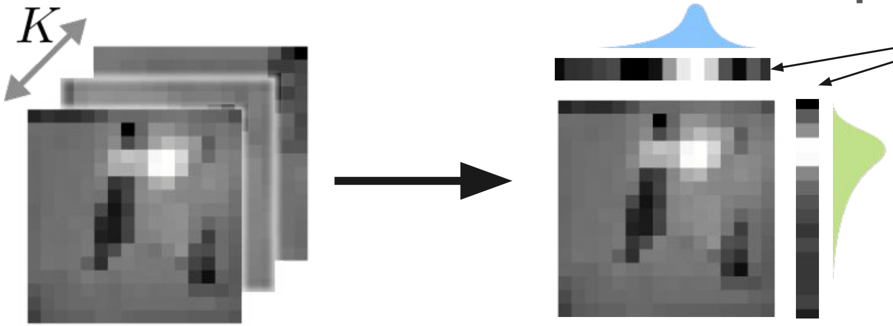
for each heatmap

2 vectors

heatmap for each keypoint

row-wise & column-wise sum

Bottleneck



heatmap for each keypoint

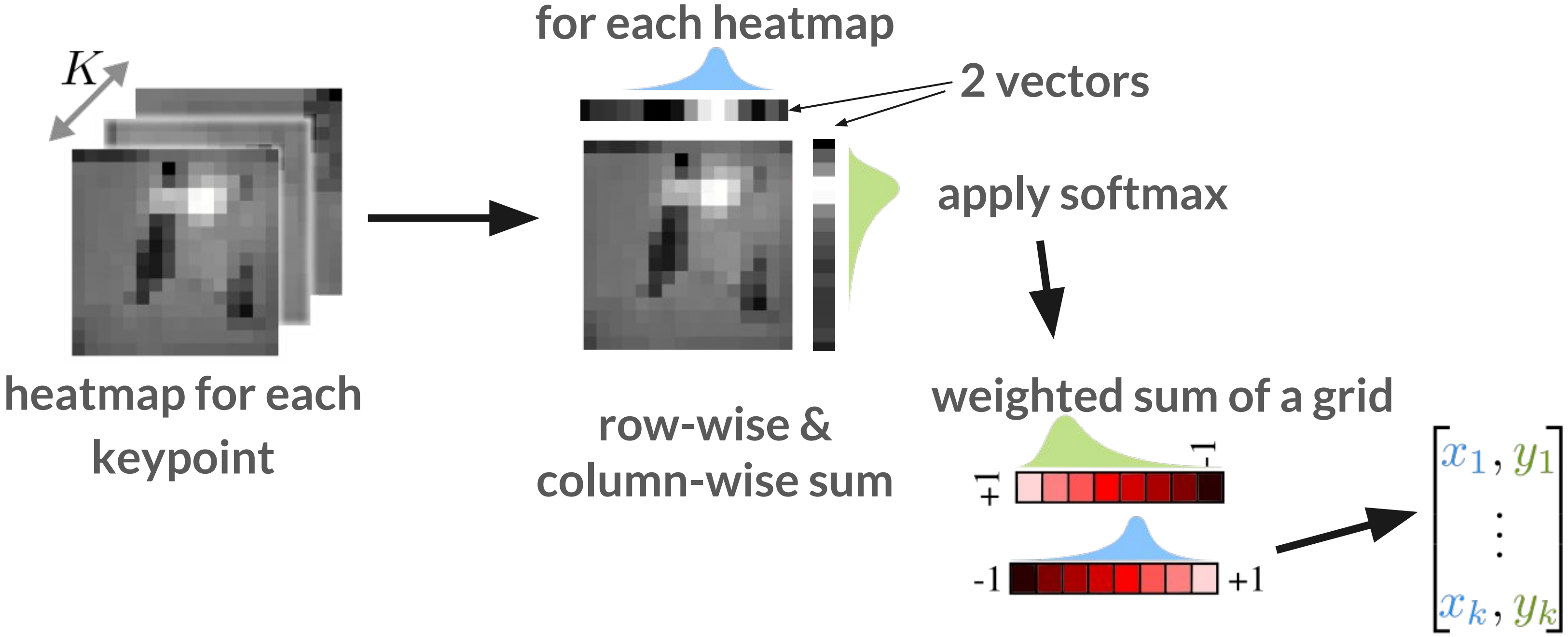
for each heatmap

2 vectors

apply softmax

row-wise & column-wise sum

Bottleneck



heatmap for each keypoint

for each heatmap

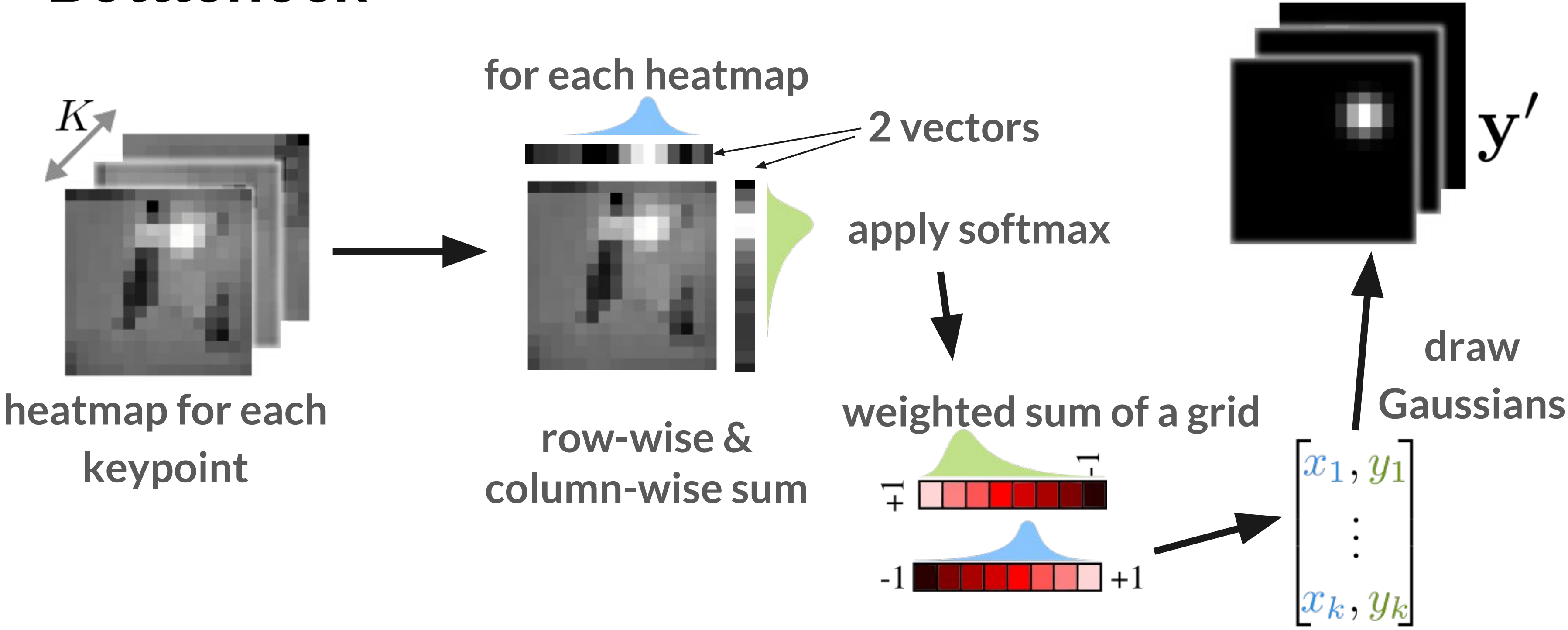
row-wise & column-wise sum

2 vectors

apply softmax

weighted sum of a grid

Bottleneck



heatmap for each keypoint

for each heatmap

2 vectors

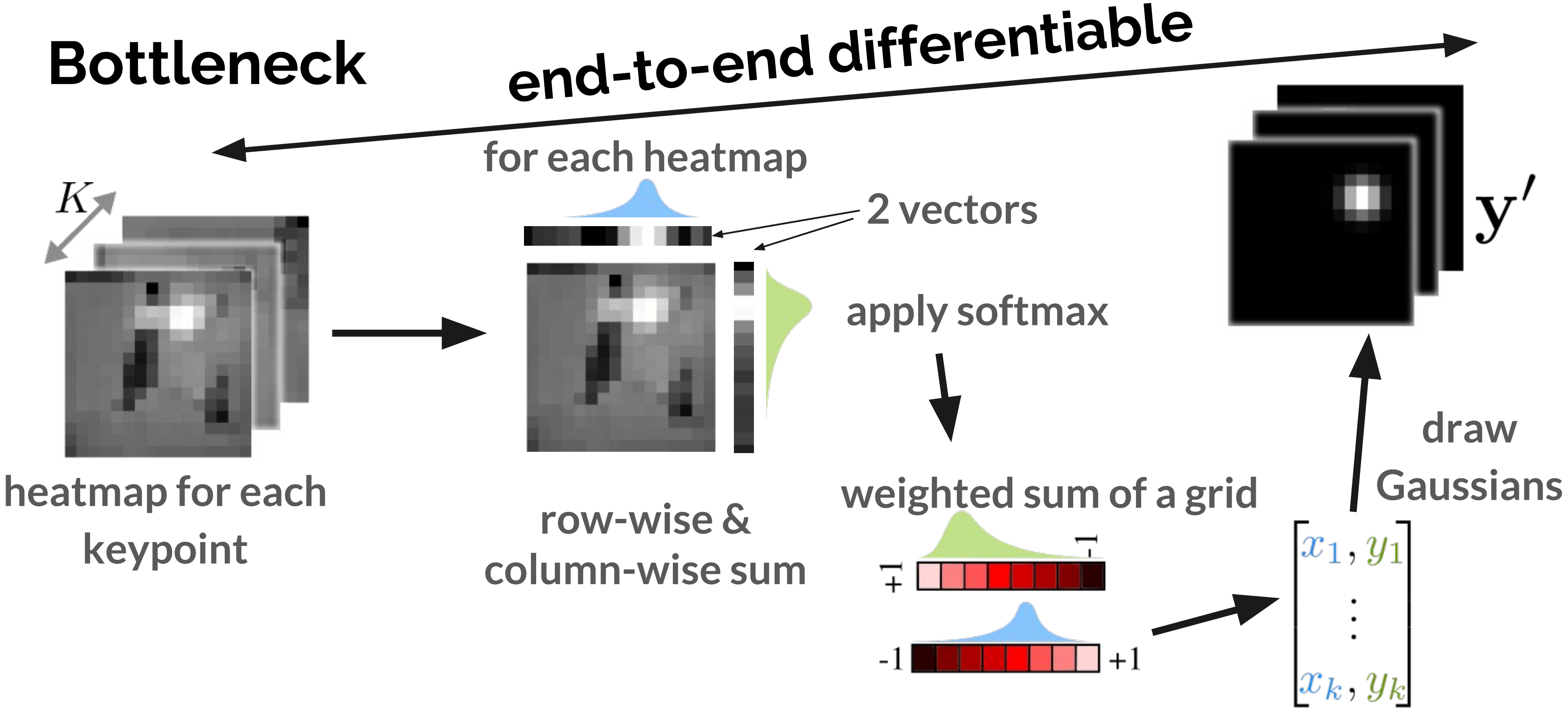
row-wise & column-wise sum

apply softmax

weighted sum of a grid

draw Gaussians

Bottleneck



end-to-end differentiable

for each heatmap

2 vectors

heatmap for each keypoint

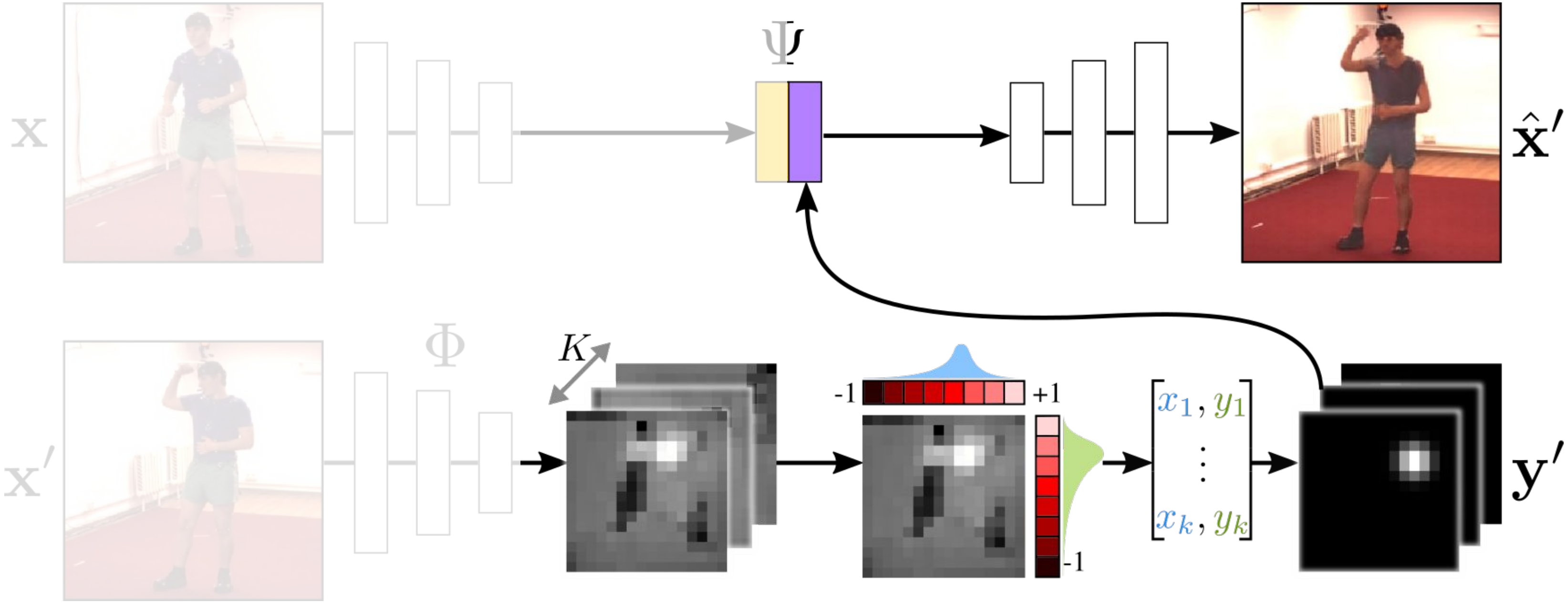
row-wise & column-wise sum

apply softmax

weighted sum of a grid

draw Gaussians

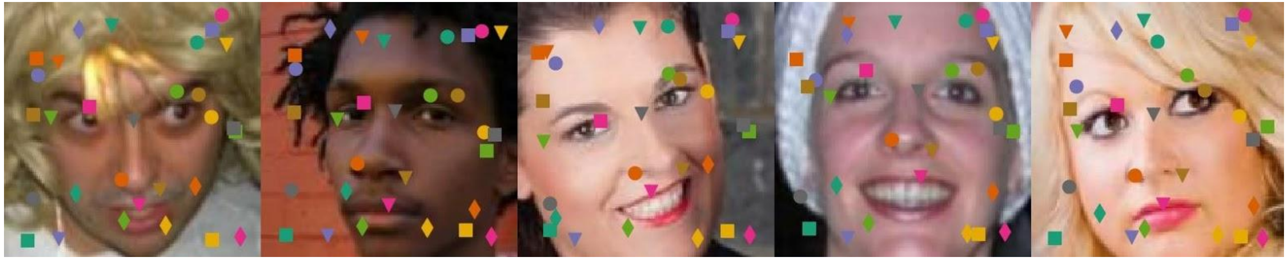
Bottleneck



$\mathbf{x}$
Ψ
$\hat{\mathbf{x}}'$
$\mathbf{x}'$
Φ
K
-1
+1
-1
x_1, y_1
$\vdots$
x_k, y_k
$\mathbf{y}'$

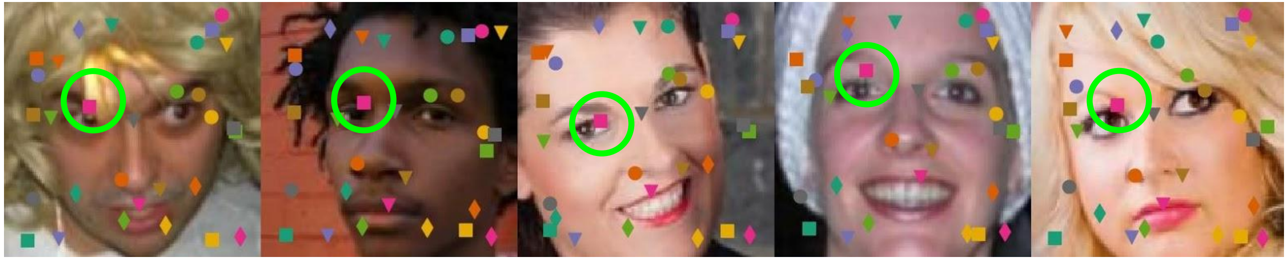
Results

Human faces



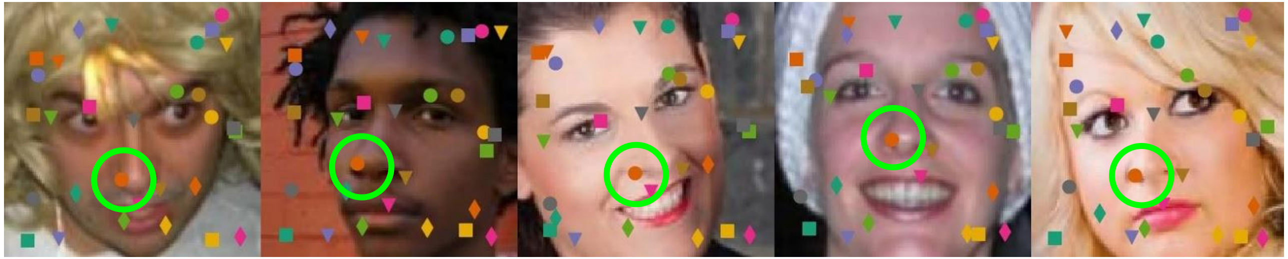
unsupervised landmarks

Human faces



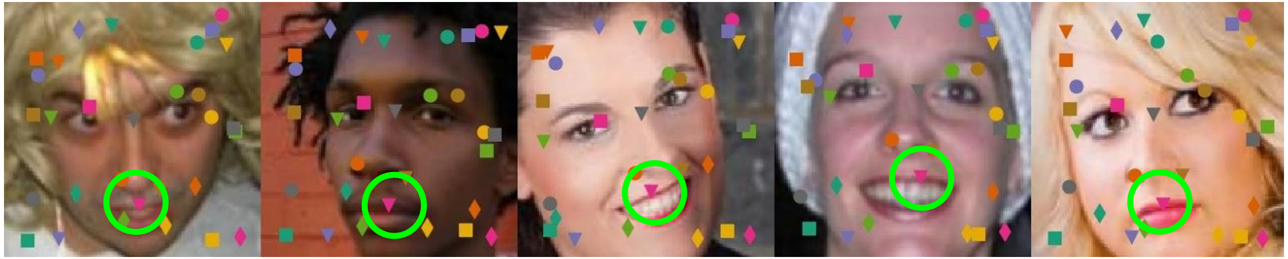
unsupervised
landmarks

Human faces



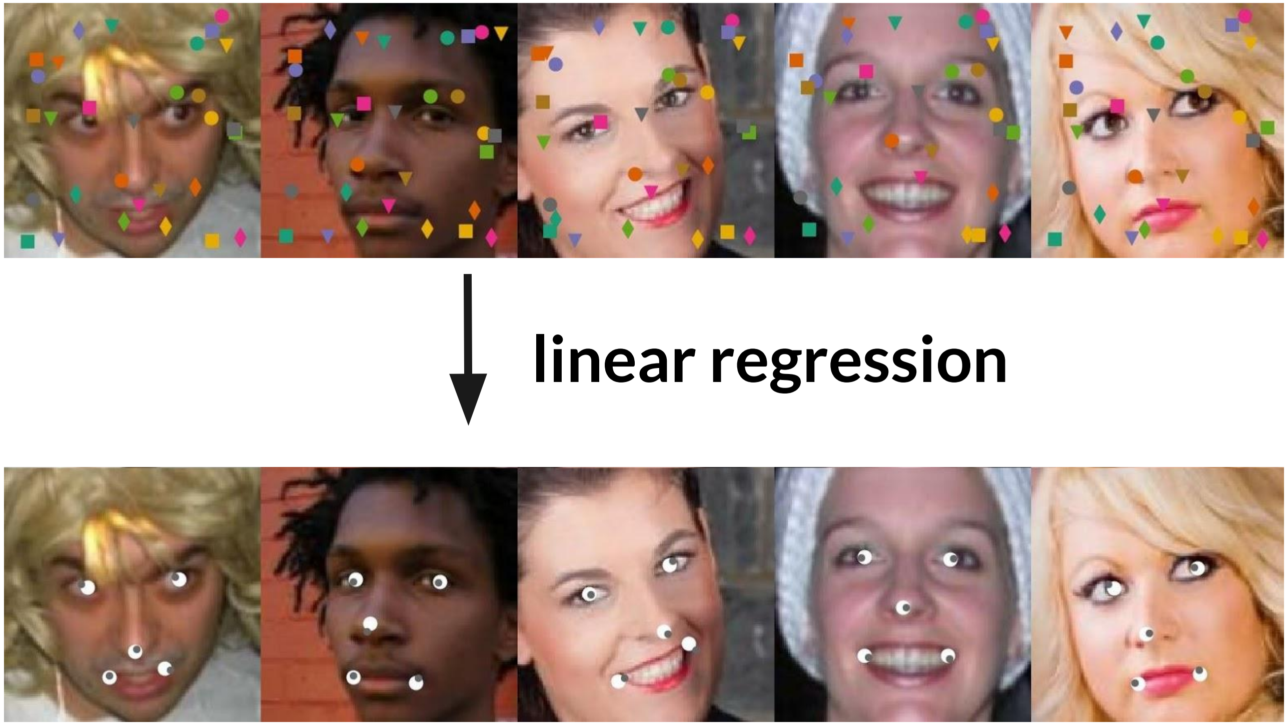
unsupervised landmarks

Human faces



unsupervised landmarks

Human faces

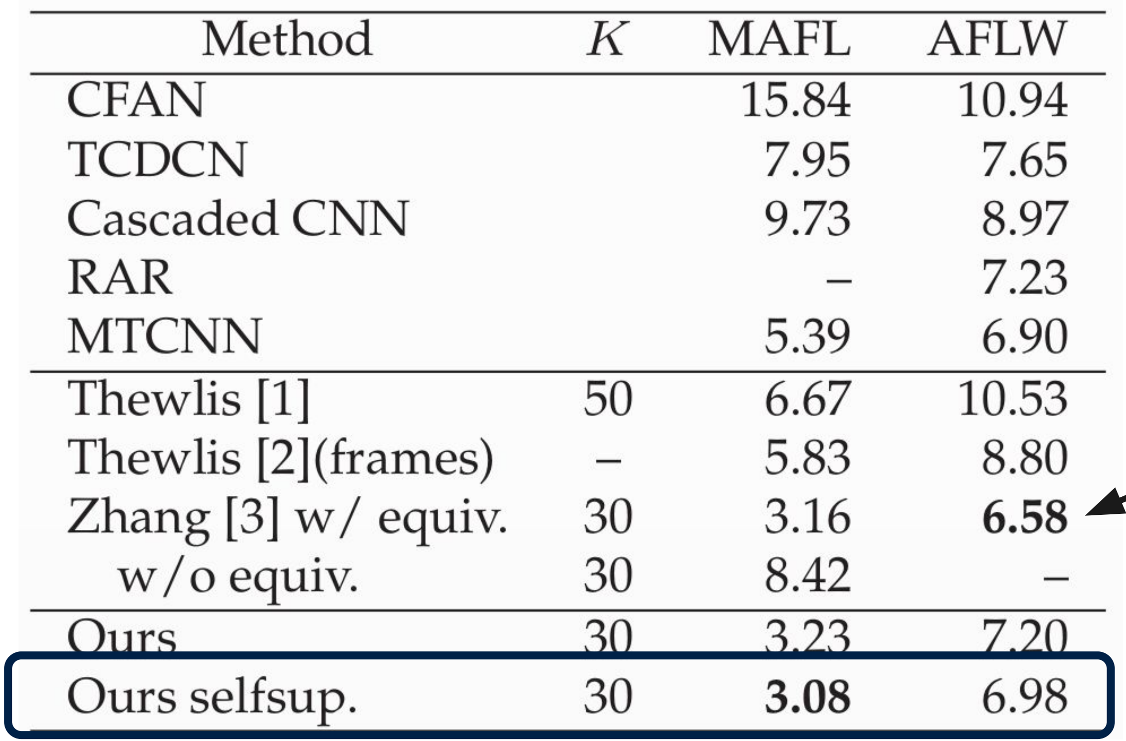


unsupervised landmarks

linear regression

regressed landmarks

Human faces



Method	K	MAFL	AFLW
CFAN		15.84	10.94
TCDCN		7.95	7.65
Cascaded CNN		9.73	8.97
RAR		–	7.23
MTCNN		5.39	6.90
Thewlis [1]	50	6.67	10.53
Thewlis [2](frames)	–	5.83	8.80
Zhang [3] w/ equiv.	30	3.16	**6.58**
w/o equiv.	30	8.42	–
Ours	30	3.23	7.20
Ours selfsup.	30	**3.08**	6.98

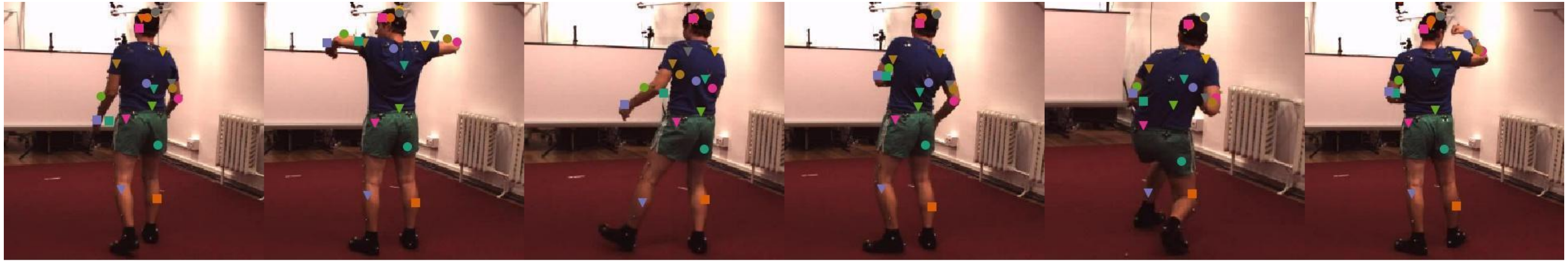
supervised methods

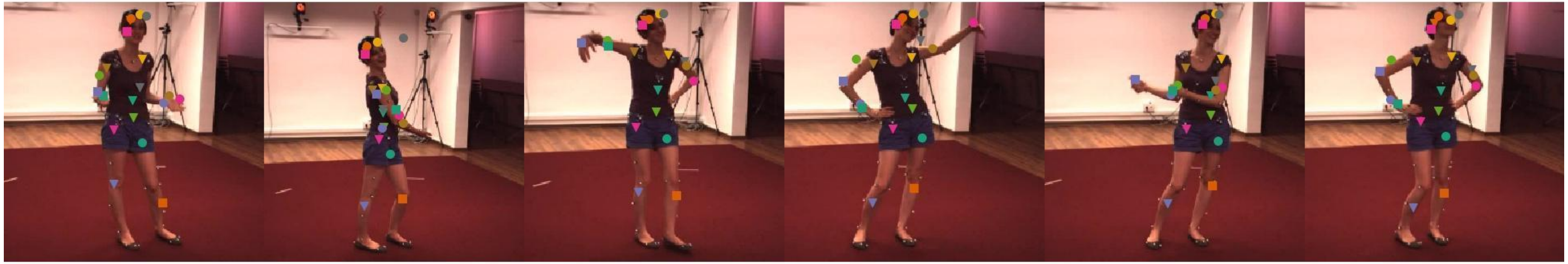
uses equivariance

unsupervised methods

Human pose

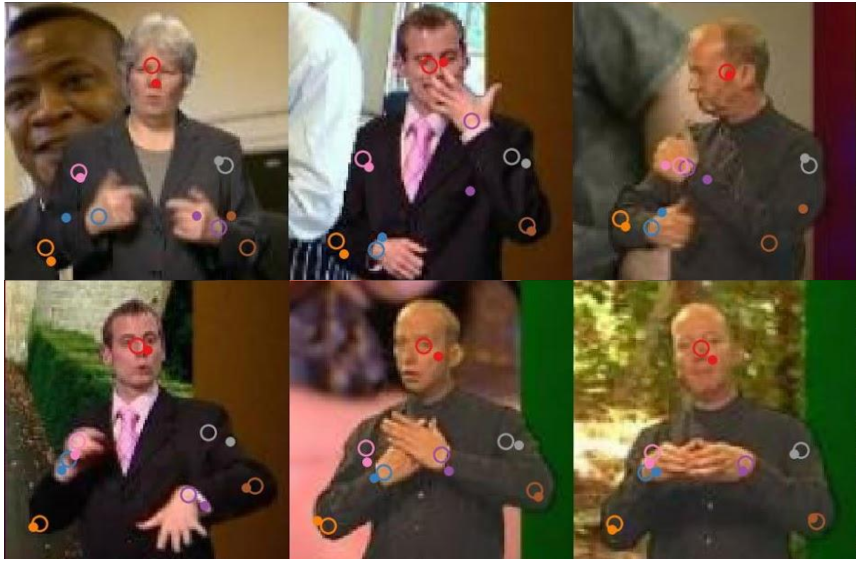
Unsupervised landmarks on Human3.6m

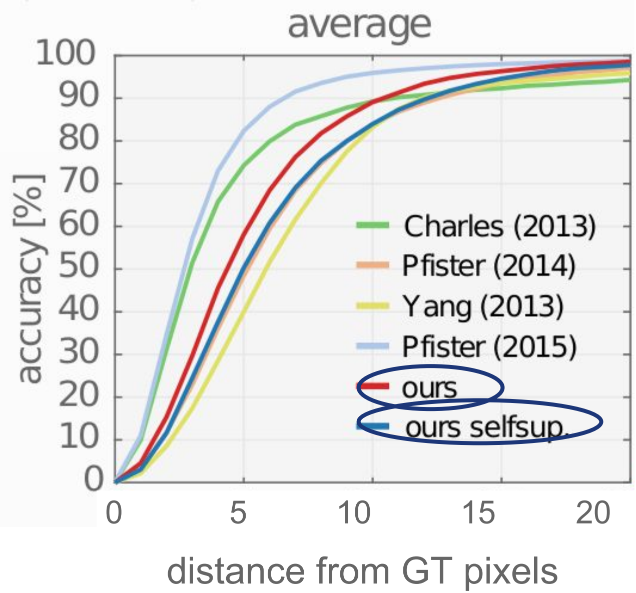




Human pose

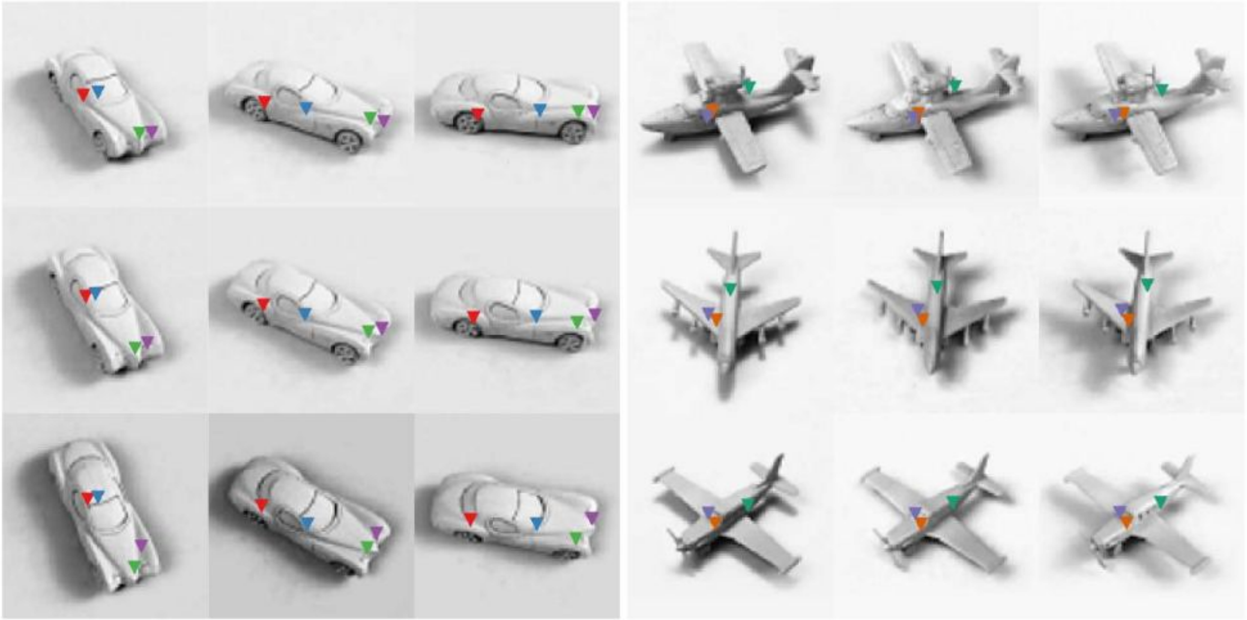
Regressed landmarks on BBCPose





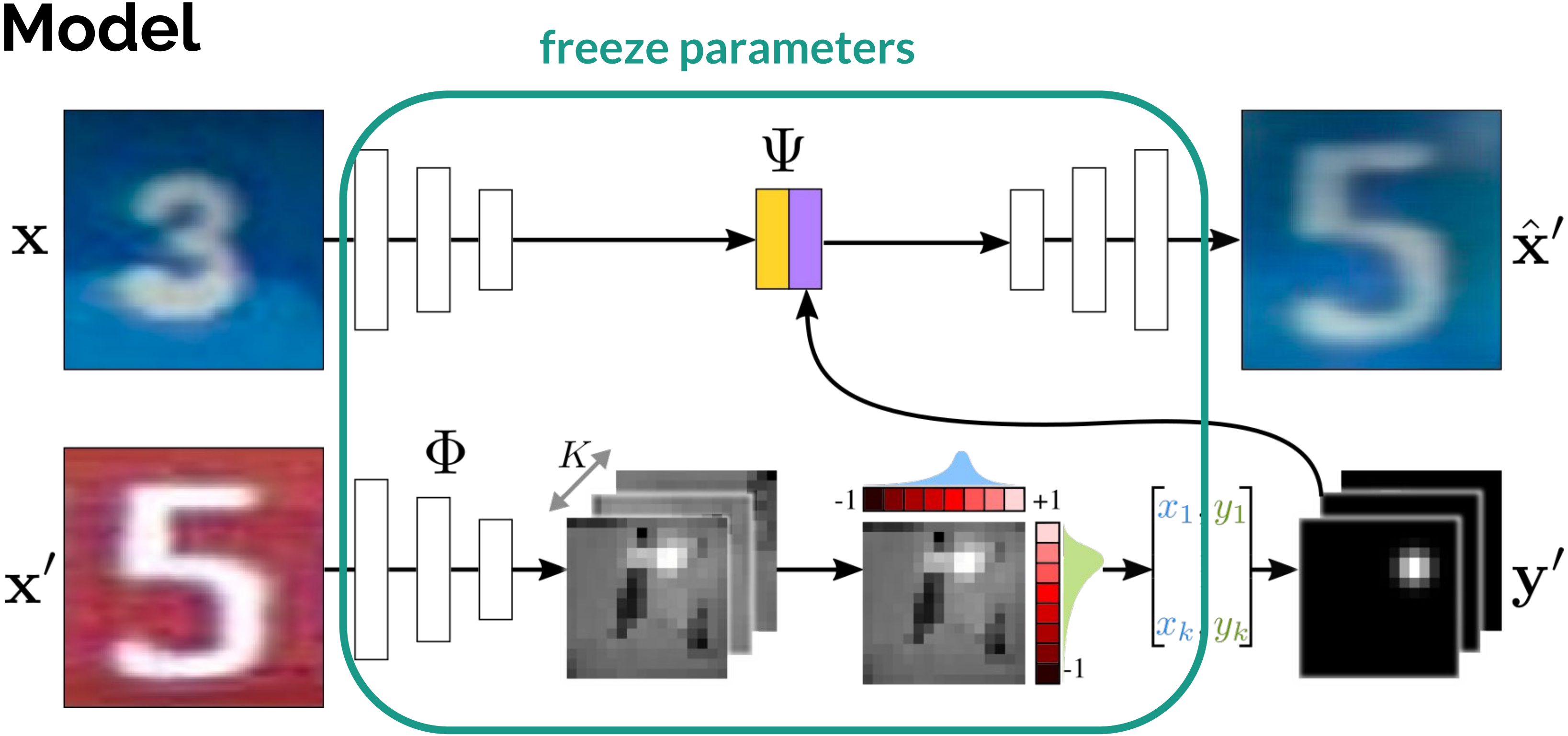
3D objects smallNORB

invariance to 3D pose, lighting and object shape



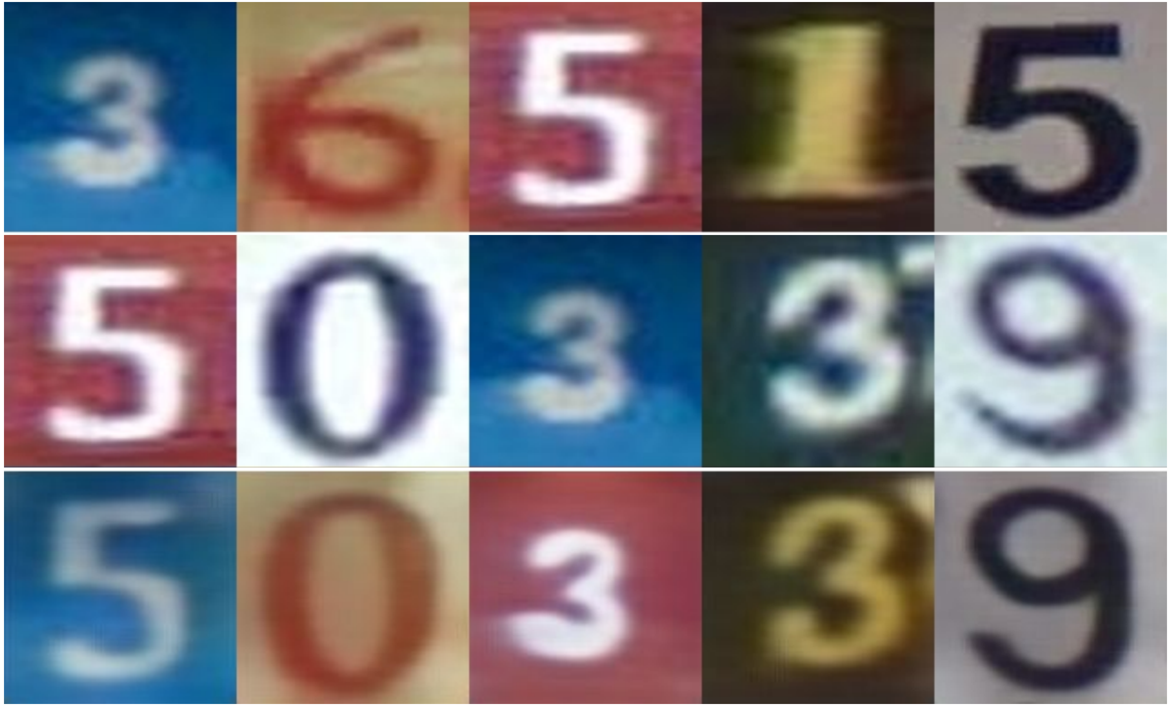
Disentangling style and geometry

Model



freeze parameters

Street numbers

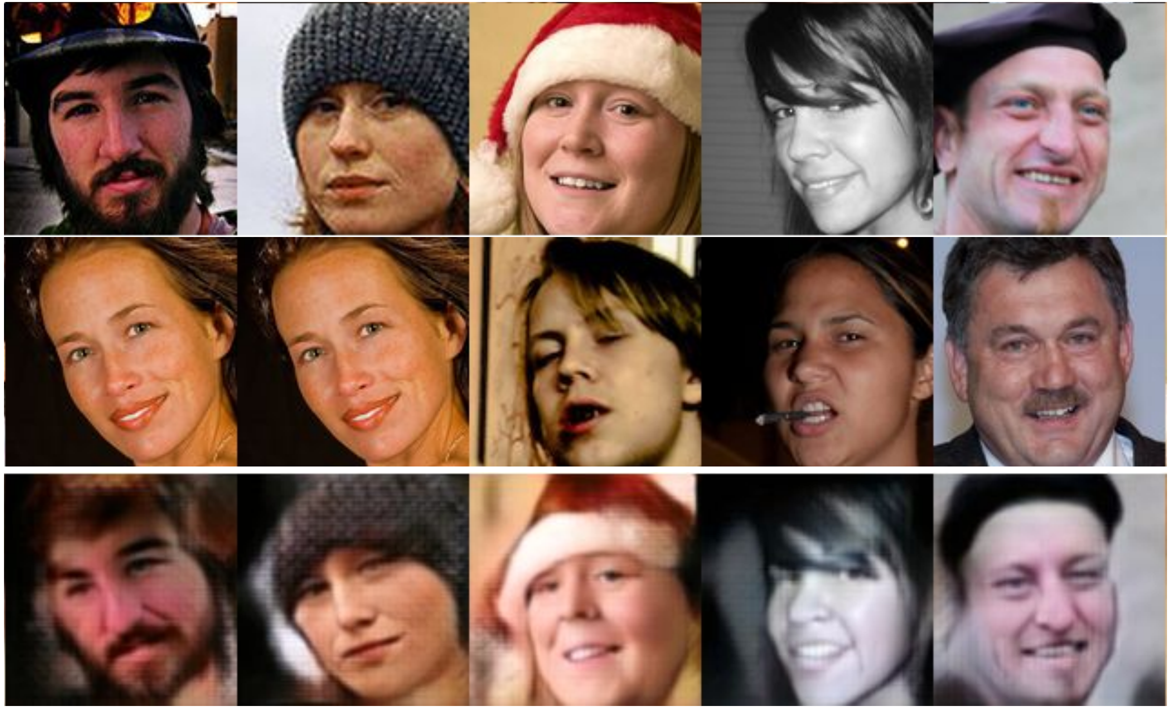


appearance

geometry

reconstruction

Human faces

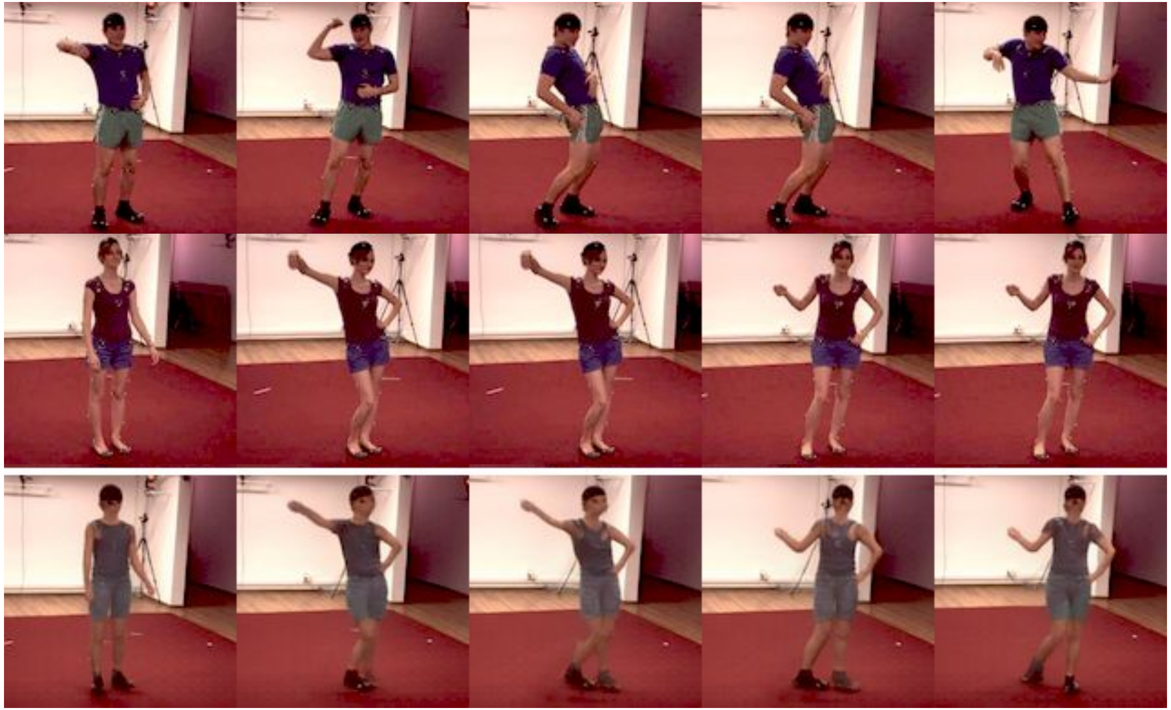


appearance

geometry

reconstruction

Human pose



appearance

geometry

reconstruction

Related work

J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In Proc. ICCV, 2017.

J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised object learning from dense invariant image labelling. In Proc. NIPS, 2017.

Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. In Proc. CVPR, 2018.

C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In Proc. NIPS, pages 613–621, 2016.

DeTone, Daniel, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description." arXiv preprint arXiv:1712.07629 (2017).

Hu, Q., Szabó, A., Portenier, T., Zwicker, M., & Favaro, P. (2017). Disentangling Factors of Variation by Mixing Them. arXiv preprint arXiv:1711.07410.

Denton, E. L. (2017). Unsupervised learning of disentangled representations from video. In Advances in Neural Information Processing Systems (pp. 4414-4423).

Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., & Murphy, K. (2018). Tracking Emerges by Colorizing Videos. arXiv preprint arXiv:1806.09594.

Conditional Image Generation for Learning the Structure of Visual Objects

Tomas Jakab*[1], Ankush Gupta*[1], Hakan Bilen[2], Andrea Vedaldi[1]

[1] VGG, University of Oxford, [2] University of Edinburgh

* equal contribution

arxiv.org/abs/1806.07823