Unsupervised Discovery of Object Landmarks as Structural Representations

Yuting Zhang †, Yijie Guo †, Yixin Jin †, Yijun Luo †, Zhiyuan He †, Honglak Lee ‡†

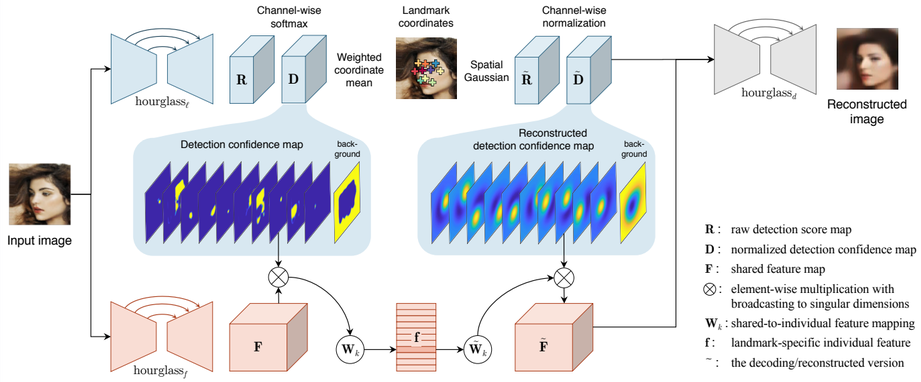
University of Michigan

CVPR 2018 SALT LAKE CITY • JUNE 18-22

1 Unsupervised Learning of Explicit Representation

- Deep neural networks can model images with rich latent representations but cannot naturally conceptualize structures of object.
- Our method **learns object structures as explicit representations** for image modeling. In particular, we proposed an autoencoder framework to **discover meaningful object landmarks without supervision**.
- Landmarks: Salient points semantically consistent across object instances.

2 Landmarks as Intermediate Representations in Autoencoder



3 Desirable Properties of Object Landmarks as Regularization

- Heatmap concentration constraint: reduce the spatial variance of $\mathbf{D}$

$$L_{\text{conc}} = \frac{\pi e}{2}\left(\sigma^2_{\text{det},u} + \sigma^2_{\text{det},v}\right)^2$$

- Landmark separation constraint: Landmarks cannot be too close

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

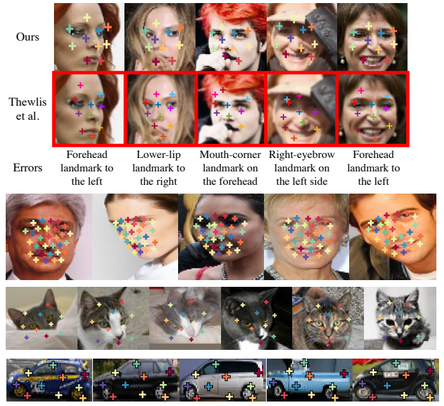
- Equivariance constraint: when the image warps, the landmark coordinates should warp accordingly. Random thin-plate-spline (TPS) warping is used in training.

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

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

4 Landmark Discovery Results

- Our method can discover visually meaningful and stable landmarks on several object categories.



5 Human-annotated Landmark Prediction

- Train a linear regressor to map the discovered landmarks to the human annotated landmarks.

Human facial landmark detection

- Outperforming previous unsupervised methods and off-the-shelf fully supervised methods.

	Method	MAFL	ALFW
Fully supervised	TCDCN	7.95	7.65
	Cascaded CNN	9.73	8.97
	RAR	-	7.23
	MTCNN	5.39	6.90
Unsupervised discovery	Thewlis et al. (50 landmarks)	6.67	10.53
	Thewlis et al. (dense frames)	5.83	8.80
	Ours (10 landmarks)	**3.46**	**7.01**
	Ours (30 landmarks)	**3.15**	**6.58**

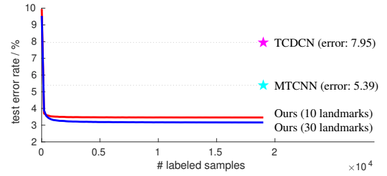
Ablative study on CelebA/MAFL

- All training objectives contribute to the final results.

Full model	No autoencoder	No concentration	No equivariance	No separation
3.15	3.45	3.91	8.42	16.56

Less labeled data for the linear regressor

- Useful for semi-supervised landmark detection
- < 500 labeled samples are enough



Landmark detection on other types of objects

- Wide applicability to many object categories.

Dataset	Car		Cat head		Human3.6M
# target landmarks	5		7		32
# discovered landmarks	10	24	10	20	16
Thewlis et al.	11.42	11.11	26.76	26.94	7.51
Ours	**5.87**	**5.80**	**15.35**	**14.84**	**4.14**

6 Attribute Prediction using Discovered Landmarks

- Our discovered landmarks are effective representations for shape-related facial attribute prediction.

Predicting 13 binary facial attributes related to the facial shape on CelebA.

Method	Feature dimension	Accuracy
Ours (discovered landmarks)	60	83.2
FaceNet (top-layer)	128	80.0
FaceNet (top-layer) + Ours	188	**84.3**
FaceNet (conv-layer)	1792	82.4
FaceNet (conv-layer) + Ours	1852	**84.1**

7 Image Manipulation using Discovered Landmarks

- Our model provides a landmark-conditioned image decoder.
- The discovered landmarks are disentangled from the appearance latent features automatically.

reconstructed manipulated