Unsupervised Part-Based of Object Shape and Appearance

CVPR2019 Oral

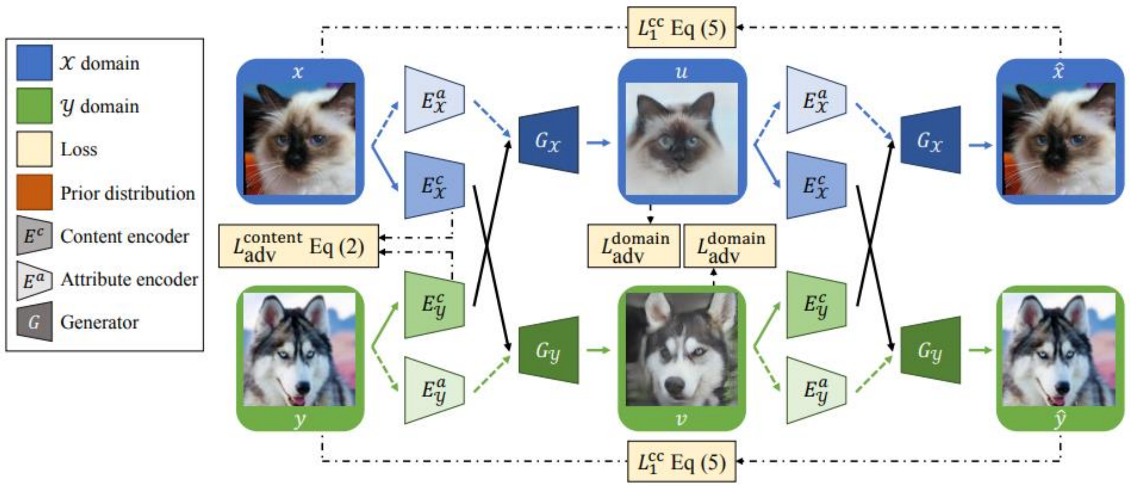
2019.08.09

발표자 박성현

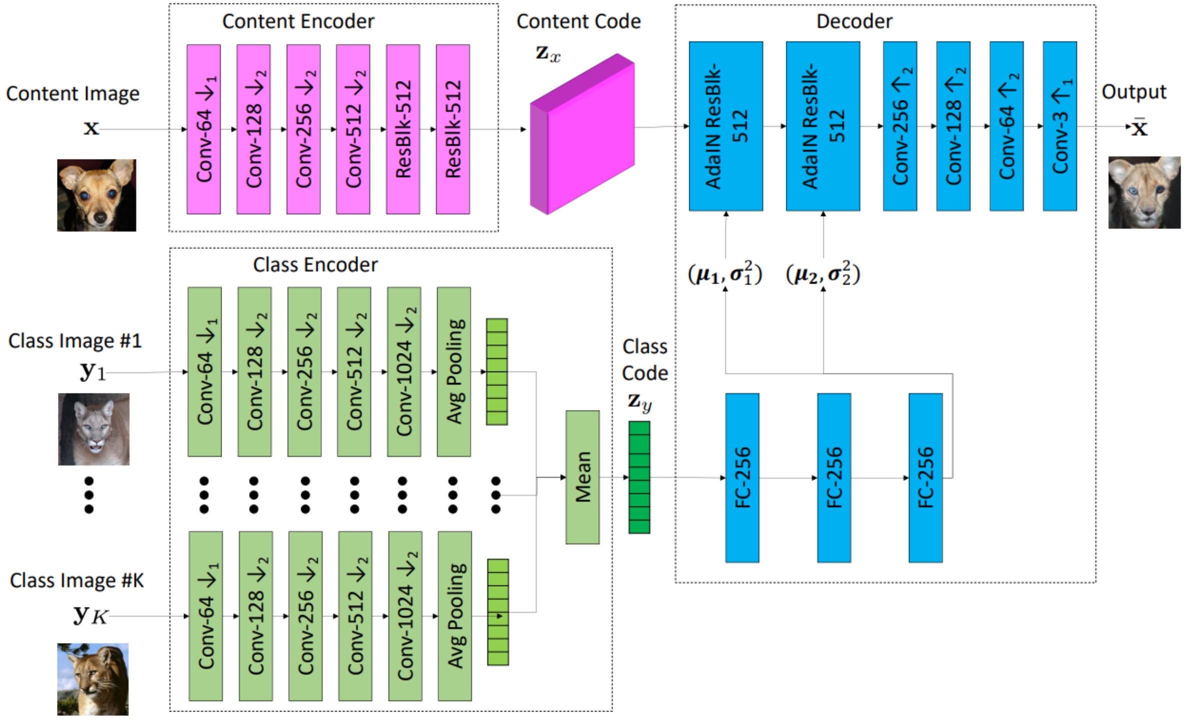
DAVIAN
Data and Visual Analytics Lab
KOREA UNIVERSITY

1 Introduction

Disentangling Shape and Appearance



[DRIT]

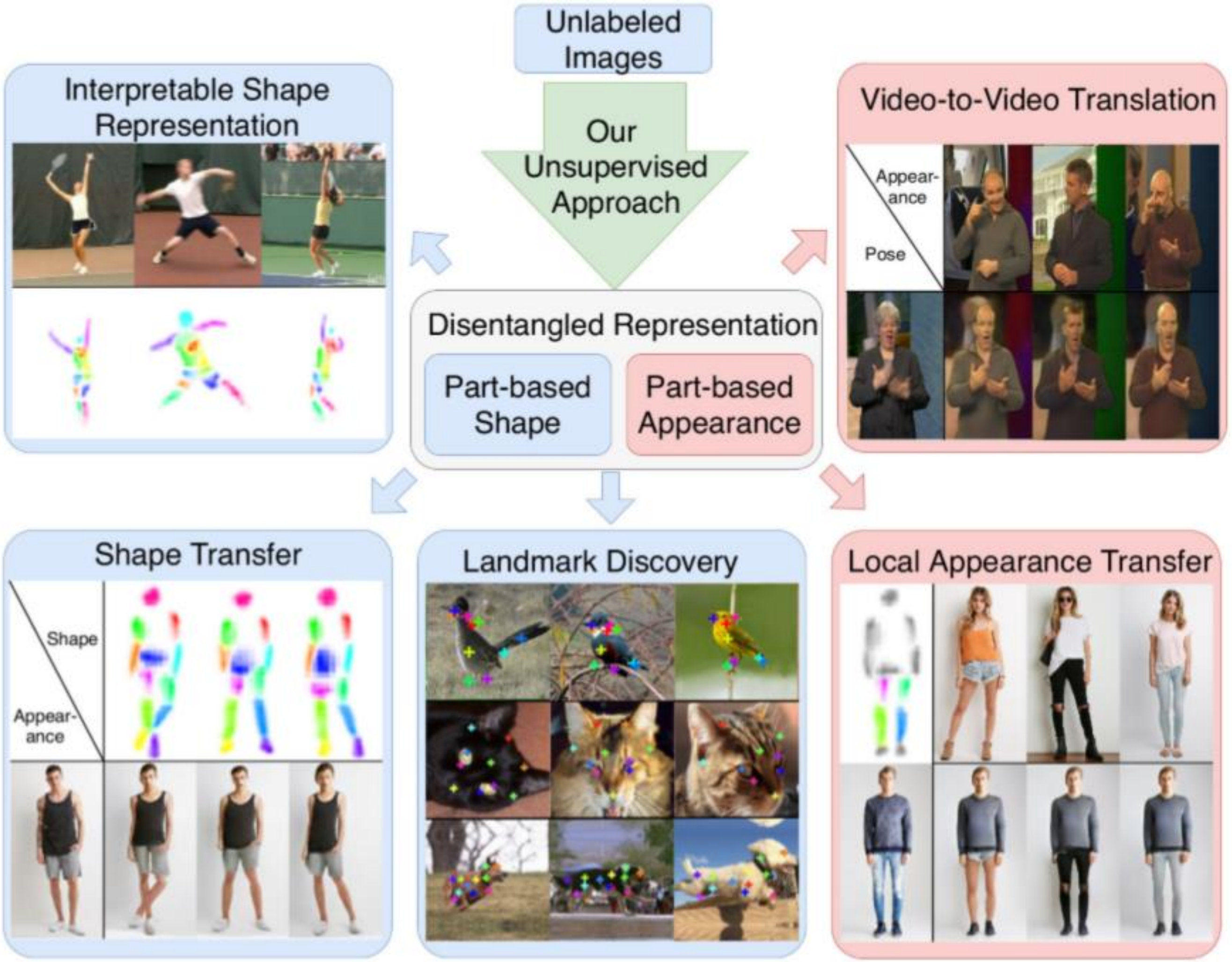


[FUNIT]

Content(Shape 정보)와 Style(Appearance 정보)로 Disentangle하는 연구가 다양하게 진행되고 있음.

1 Introduction

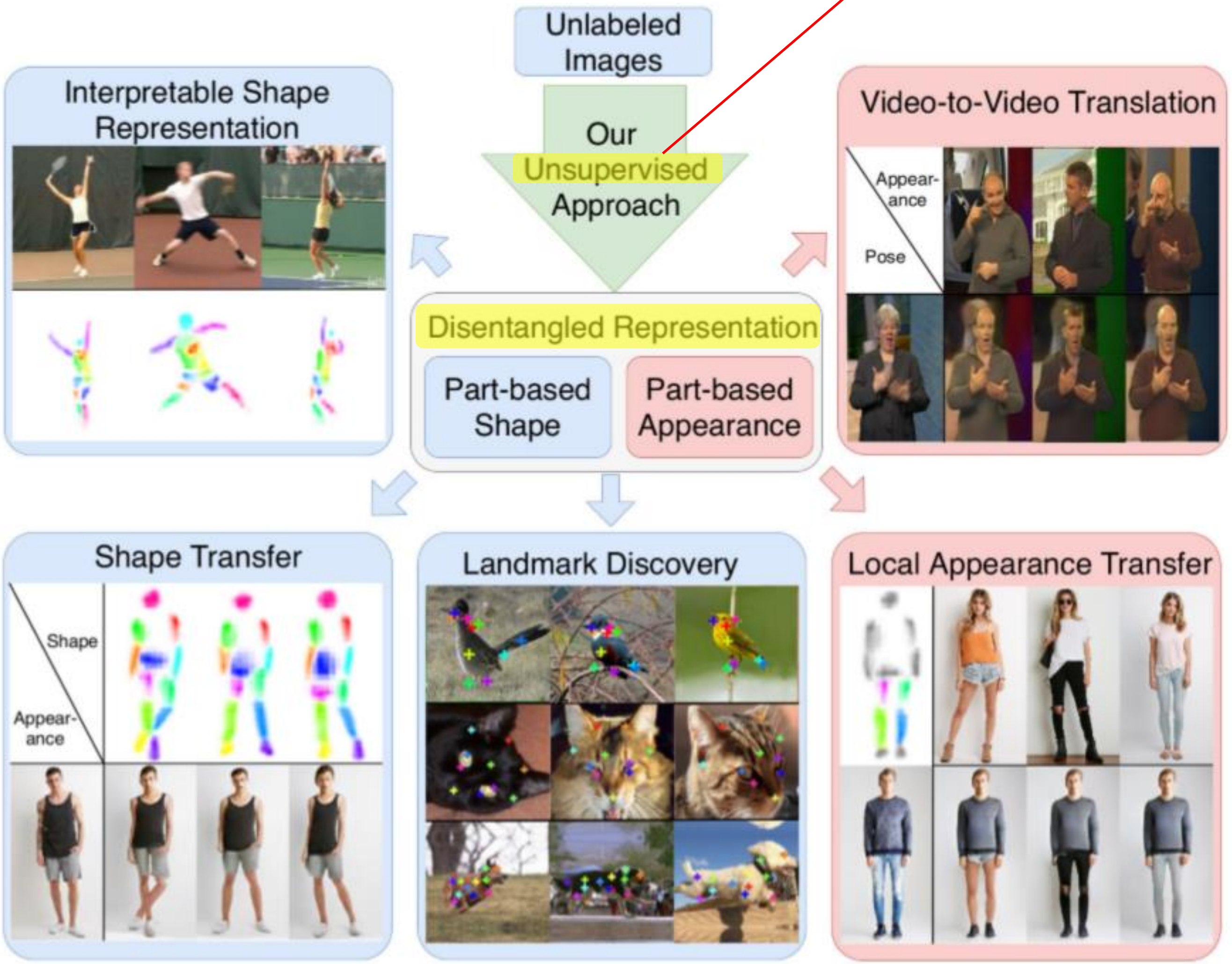
Disentangling Shape and Appearance



1 Introduction

Disentangling Shape and Appearance

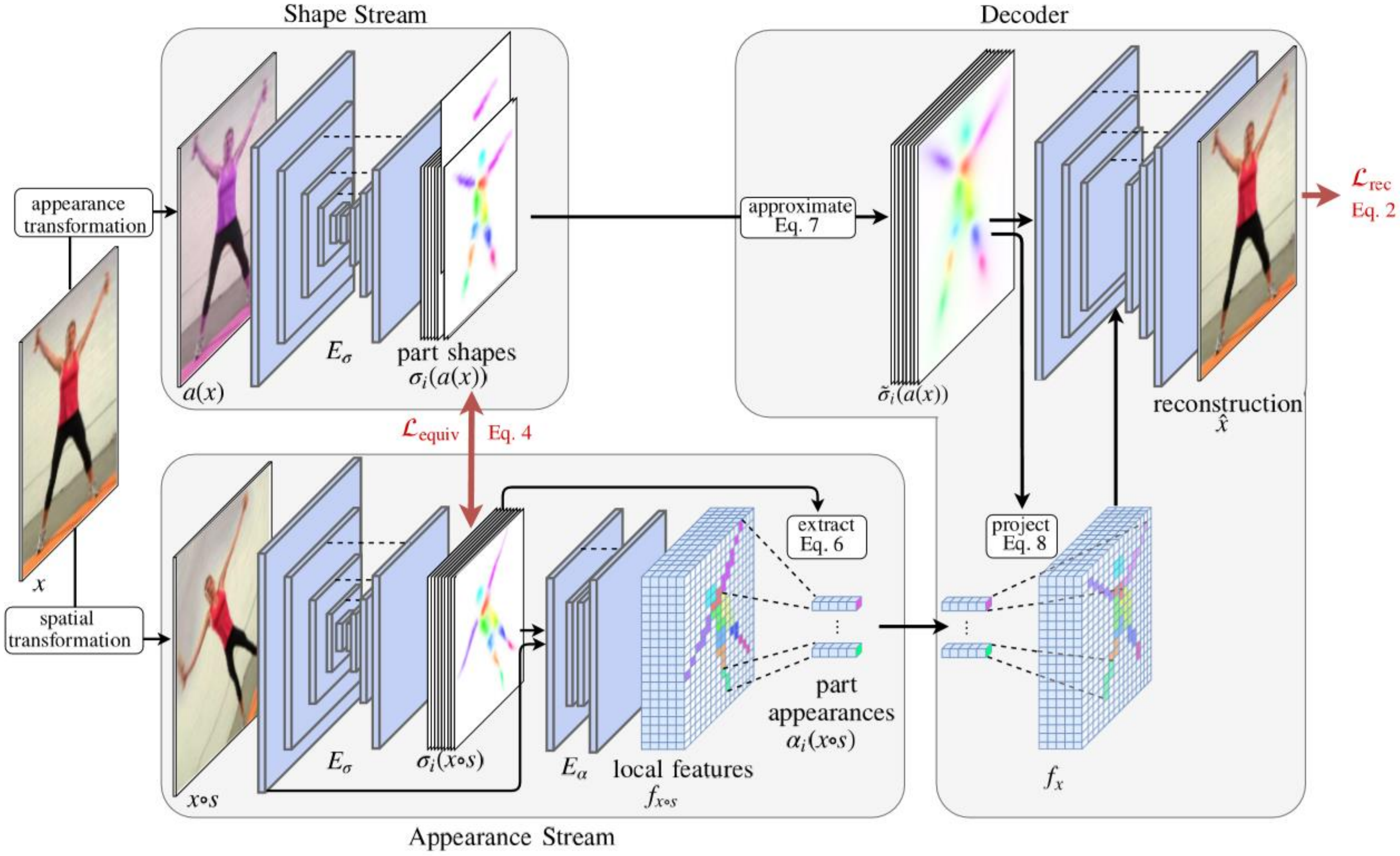
Label이 필요 X



2 Model

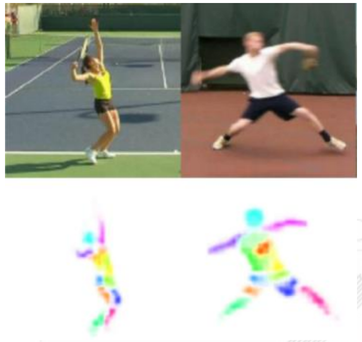
Model Overview

$$\phi_i(x) = [\alpha_i(x), \sigma_i(x)] \stackrel{!}{=} [\alpha_i(x \circ s), \sigma_i(a(x))]$$



2 Model

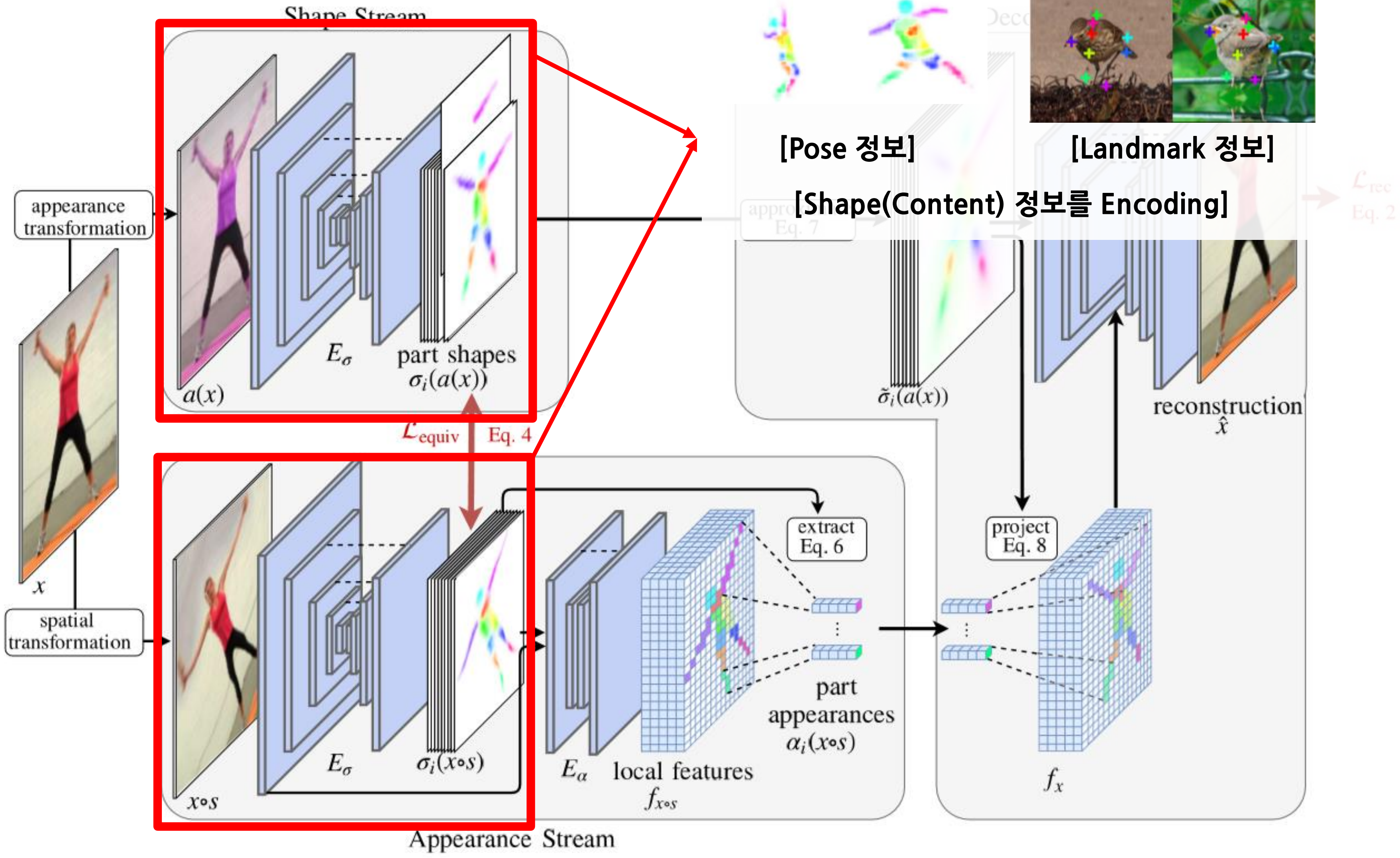
Model Overview



[Pose 정보]

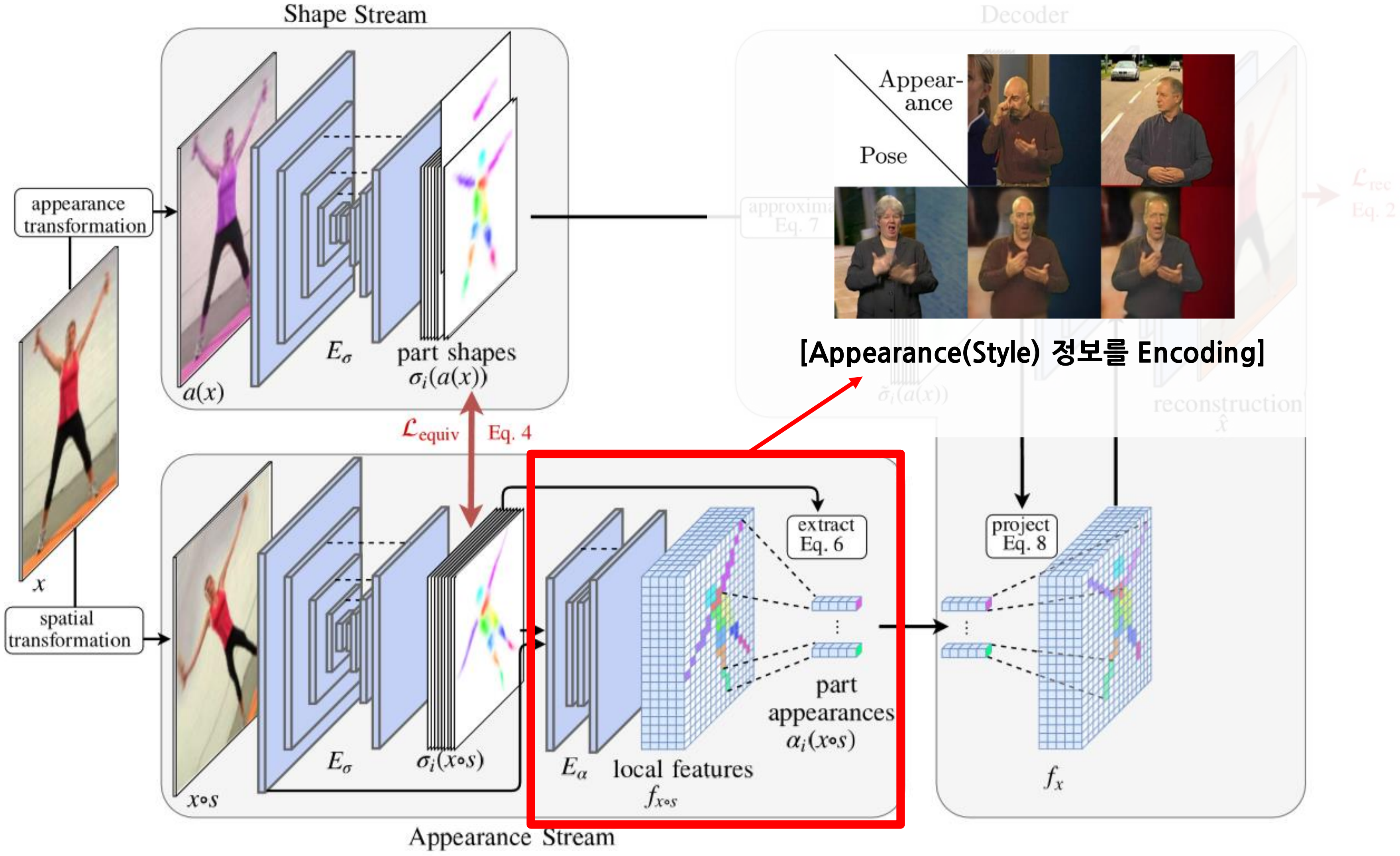
[Landmark 정보]

[Shape(Content) 정보를 Encoding]



2 Model

Model Overview

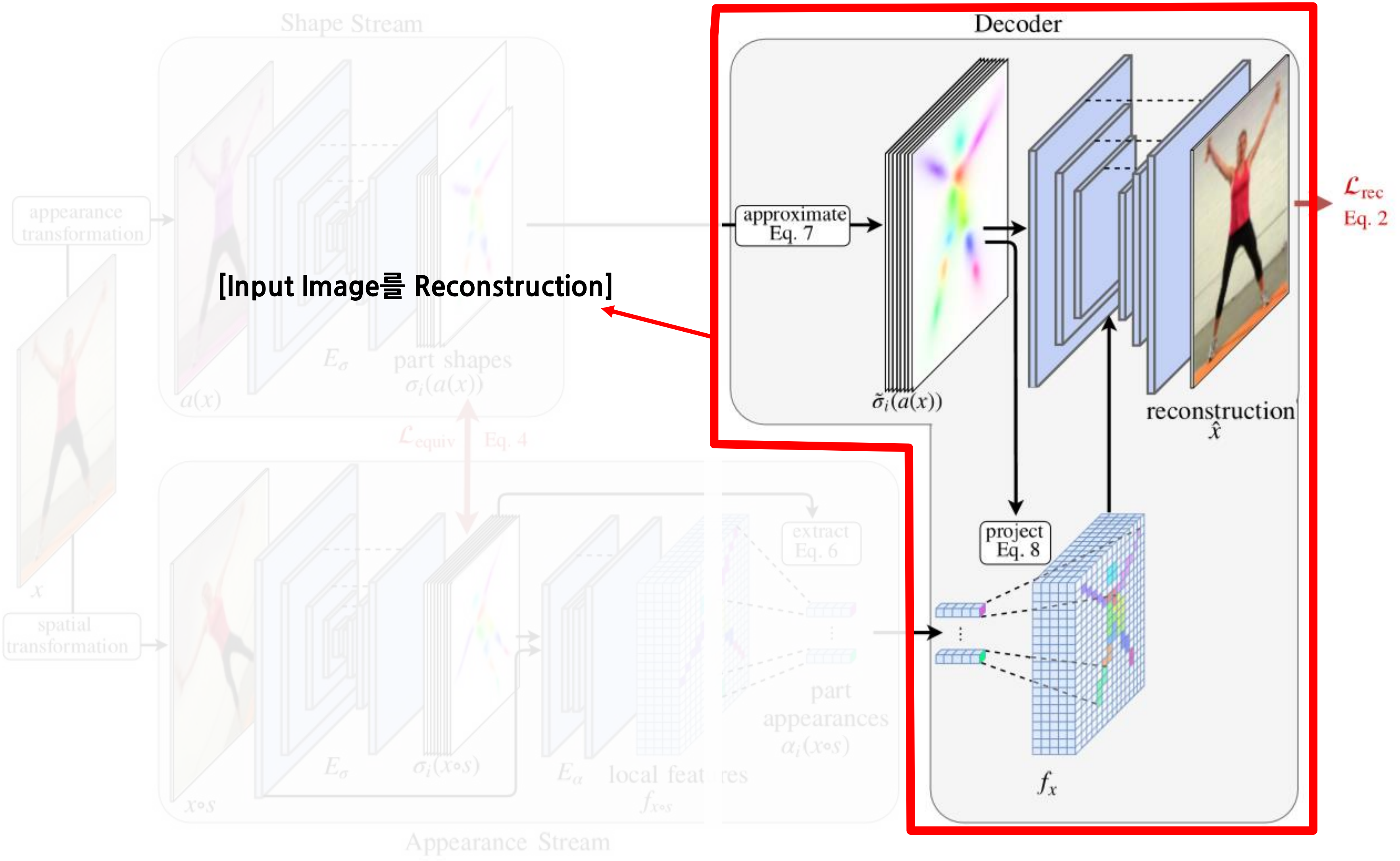


[Appearance(Style) 정보를 Encoding]

2

Model

Model Overview

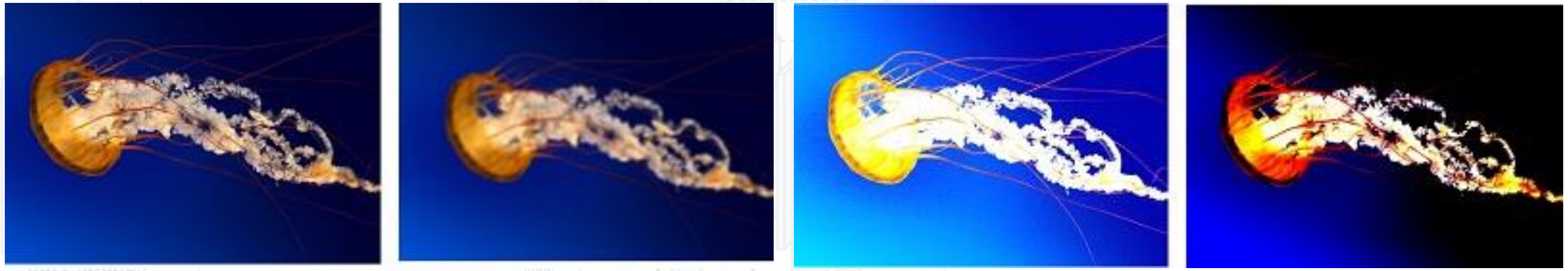


2 Model

Appearance & Spatial Transformation

- Appearance Transformation (a)

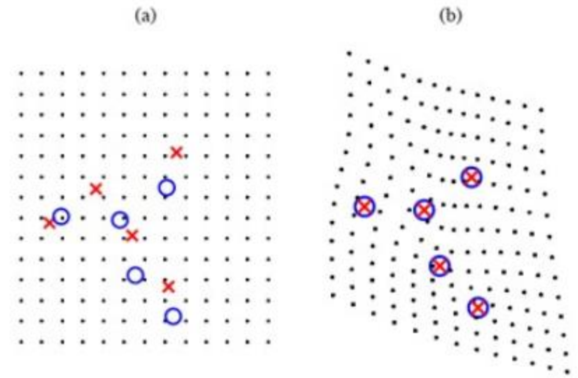
Changes in brightness, contrast, and hue

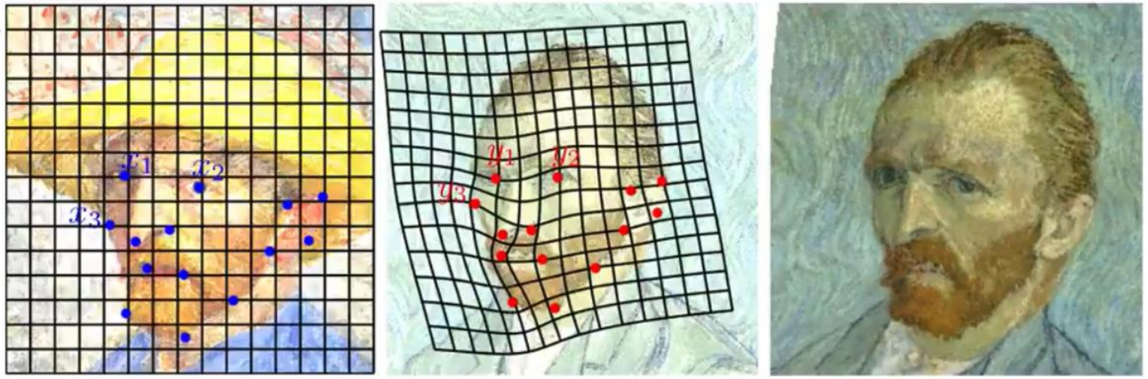


- Spatial Transformation (s)

Thin plate spline(TPS) Transformation (Image의 Shape을 변화하는 데 사용)

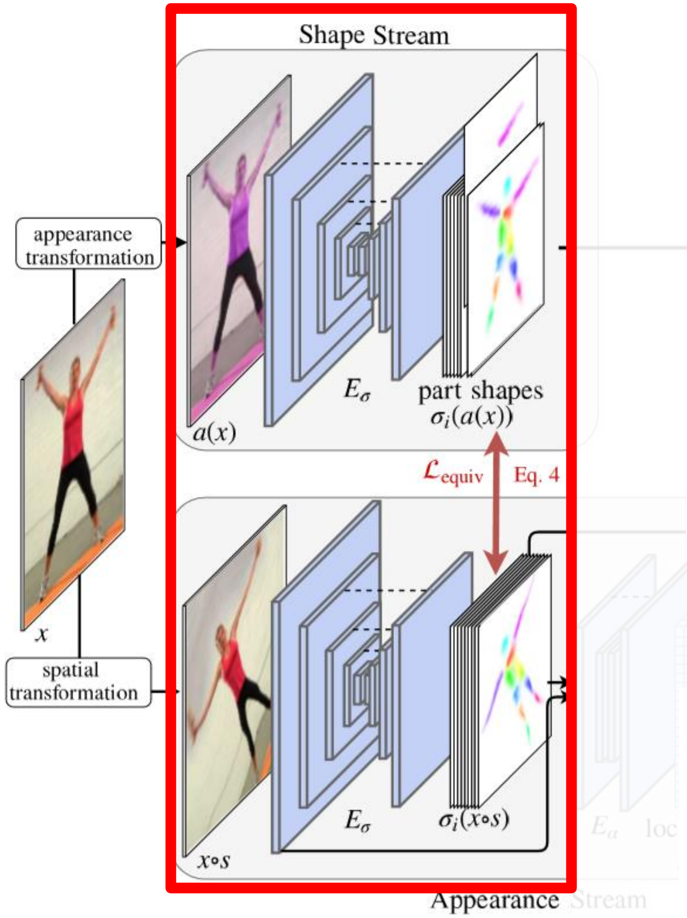
Randomly sample another frame from the same video sequence which acts as $x \circ s$





2 Model

Part Shape & Equivariance Loss



- **Part Shape Encoder**

Hourglass Network를 사용

- Equivariance Loss

Part Shape $\sigma_i(x)$는 Deformation을 해도 같아야 함.

Pixel level의 Loss를 minimize하는 방법은 실제로는 unstable했다고 함.

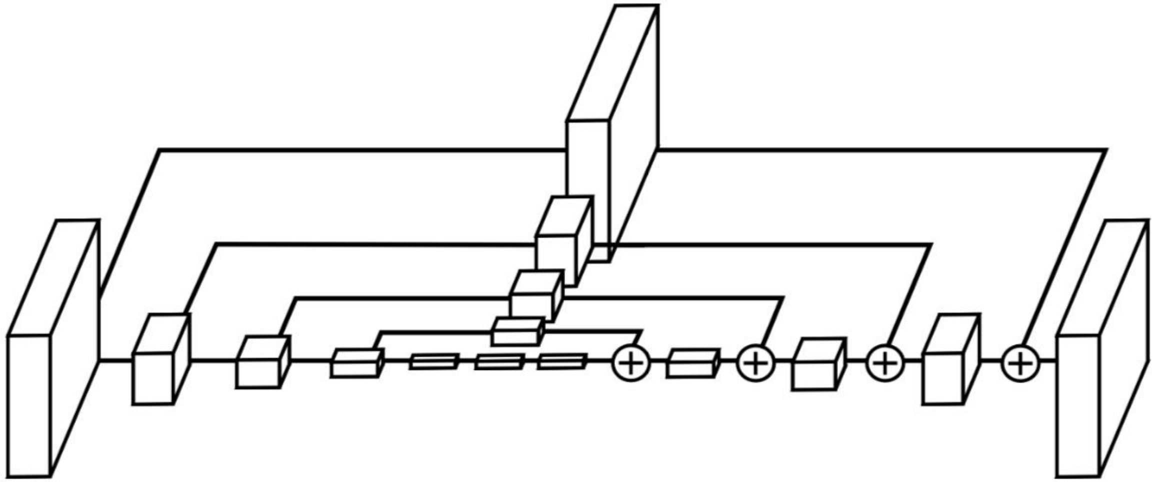
$$\sum_i \sum_{u \in \Lambda} \left\| \sigma_i(x \circ s)[u] - \sigma_i(x)[s(u)] \right\|$$

그래서 Equivariance Loss를 사용 (mean과 variance를 이용)

$$\mathcal{L}_{\text{equiv}} = \sum_i \lambda_\mu \| \mu[\sigma_i(x \circ s)] - \mu[\sigma_i(a(x)) \circ s] \|_2 + \lambda_\Sigma \| \Sigma[\sigma_i(x \circ s)] - \Sigma[\sigma_i(a(x)) \circ s] \|_1 \;,$$

2 Model

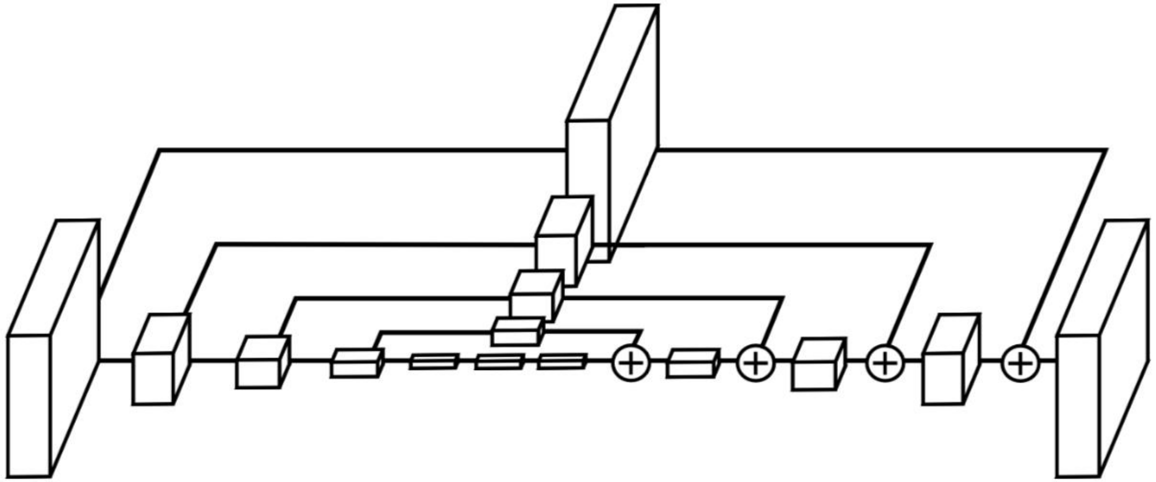
Appendix - Hourglass Network (ECCV 2016)



- **Human Pose Estimation** 분야에서 한 때 State of the Art 성능을 보였던 Model
- 얼굴이나 손과 같은 Feature들을 식별하는 것에는 Local evidence가 중요한 반면, 최종적인 Pose를 추정하기 위해서는 Full Body에 대한 정보가 필요하다. 이를 위해서는 **여러 Scale에 걸쳐 필요한 정보를 포착**해낼 수 있어야 한다.
- 해당 모델에서는 **Skip Layer를 이용**하여 Spatial Information을 유지하는 방식을 사용하였다.

2 Model

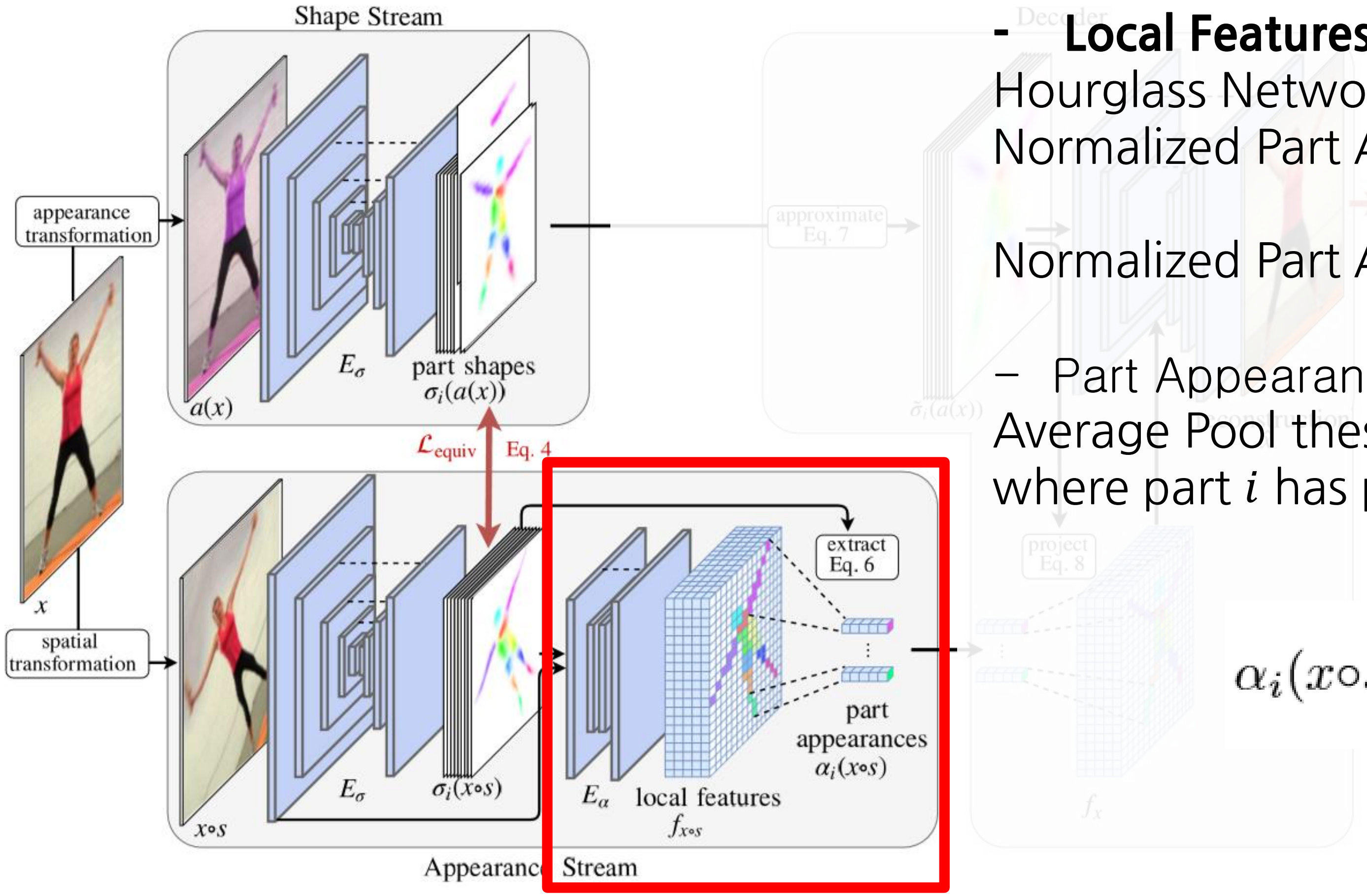
Appendix - Hourglass Network (ECCV 2016)



- Downsampling을 위해 Conv layer와 Maxpooling layer를 사용
- 매 Maxpooling 단계에서 Input을 별도의 branch로 내보내고, 이에 Conv 연산을 적용한다. 이를 통해 scale마다 feature가 추출됨
- Upsampling으로는 Nearest Neighbor Upsampling, feature와의 조합에는 Elementwise addition 연산을 이용
- 네트워크의 출력은 각 관절에 대한 Heatmap들이다.

2 Model

Part Appearance



- **Local Features Encoder**

Hourglass Network를 사용
Normalized Part Activations와 Image를 Concat해서 Input으로 사용

Normalized Part Activations :$\sigma_i(x \circ s) / \sum_{u \in \Lambda} \sigma_i(x \circ s)[u]$

- Part Appearance

Average Pool these local features at all locations where part i has positive activation

$$\alpha_i(x \circ s) = \frac{\sum_{u \in \Lambda} f_{x \circ s}[u] \sigma_i(x \circ s)[u]}{\sum_{u \in \Lambda} \sigma_i(x \circ s)[u]}$$

2 Model

Reconstructing the Original Image

- **Approximate**

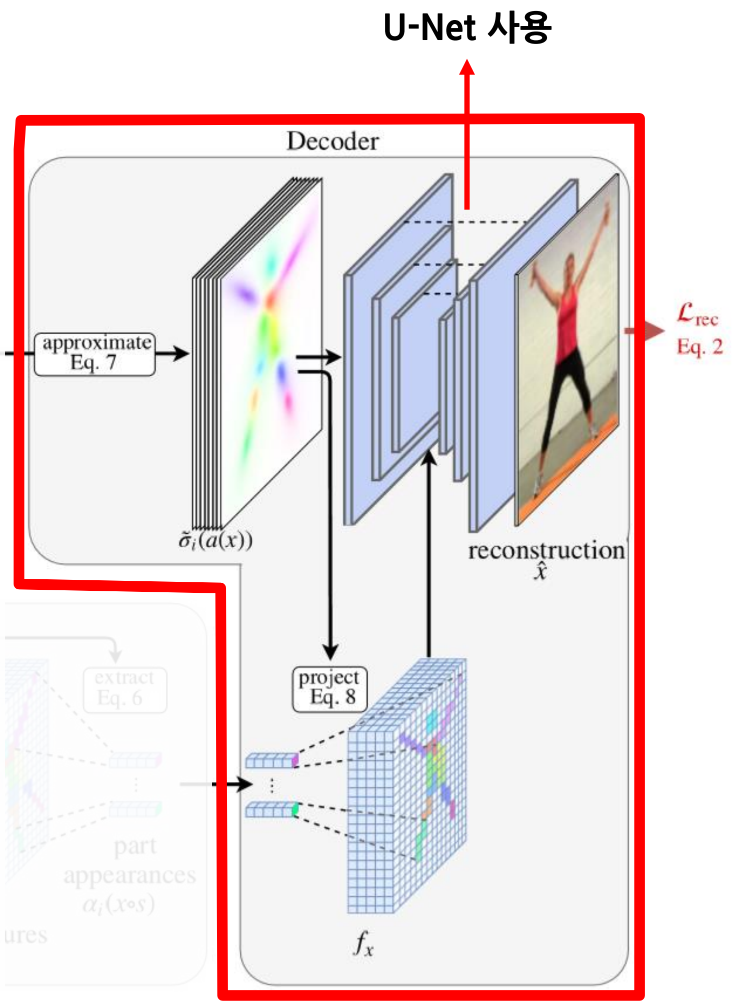
Extra information present in part activations is **neglected**, forcing the shape encoder E_σ **to concentrate on an unambiguous part localization**. (or else reconstruction loss would increase)

$$\tilde{\sigma}_i(a(x))[u] = \frac{1}{1 + (u - \mu_i)^T \Sigma_i^{-1} (u - \mu_i)}$$

- Project

The corresponding part activations $\tilde{\sigma}_i(a(x))$ designate the regions of parts i in image x to project the part appearances $\alpha_i(x \circ s)$ onto a localized appearance encoding f_x

$$f_x[u] = \sum_i \frac{\alpha_i(x \circ s) \cdot \tilde{\sigma}_i(a(x))[u]}{1 + \sum_j \tilde{\sigma}_j(a(x))[u]}$$



2 Model

Reconstructing the Original Image

- **Approximate**

Extra information present in part activations is **neglected**, forcing the shape encoder E_σ **to concentrate on an unambiguous part localization**. (or else reconstruction loss would increase)

$$\tilde{\sigma}_i(a(x))[u] = \frac{1}{1 + \boxed{(u-\mu_i)^T \Sigma_i^{-1} (u-\mu_i)}}$$

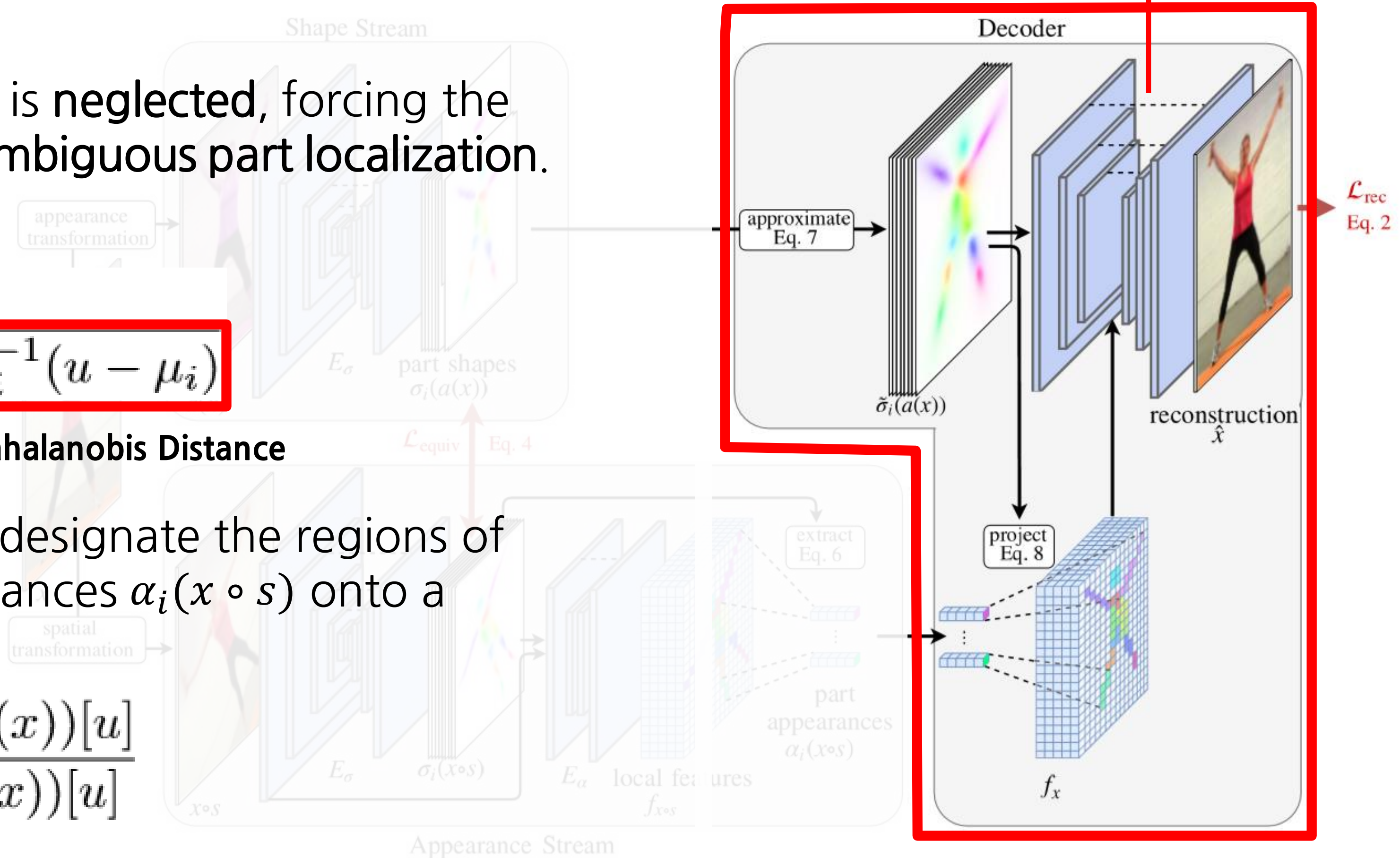
Mahalanobis Distance

- Project

The corresponding part activations $\tilde{\sigma}_i(a(x))$ designate the regions of parts i in image x to project the part appearances $\alpha_i(x \circ s)$ onto a localized appearance encoding f_x

$$f_x[u] = \sum_i \frac{\alpha_i(x \circ s) \cdot \tilde{\sigma}_i(a(x))[u]}{1 + \sum_j \tilde{\sigma}_j(a(x))[u]}$$

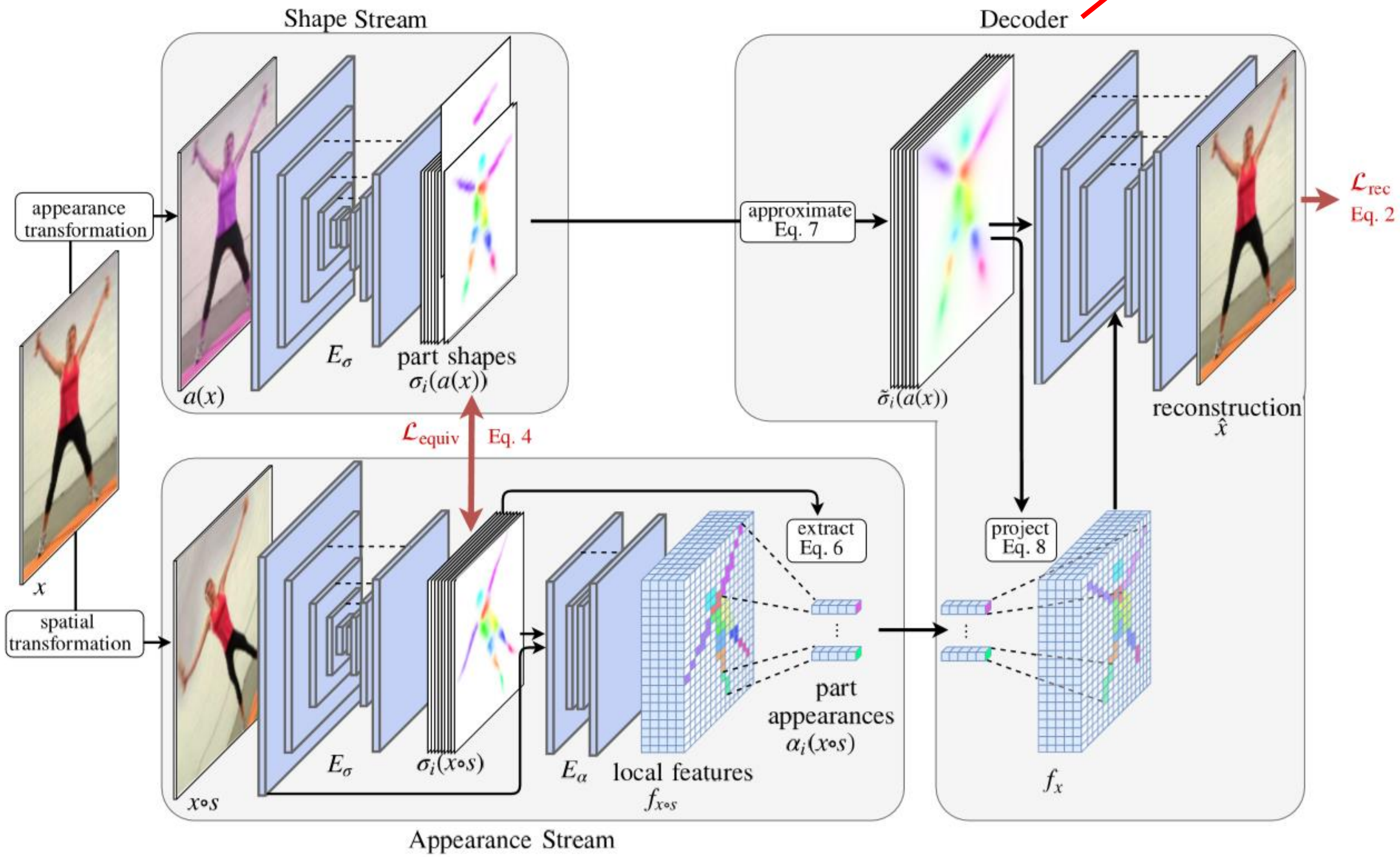
U-Net 사용



2 Model

Loss Functions

Decoder 학습에 Adversarial loss를 추가로 사용



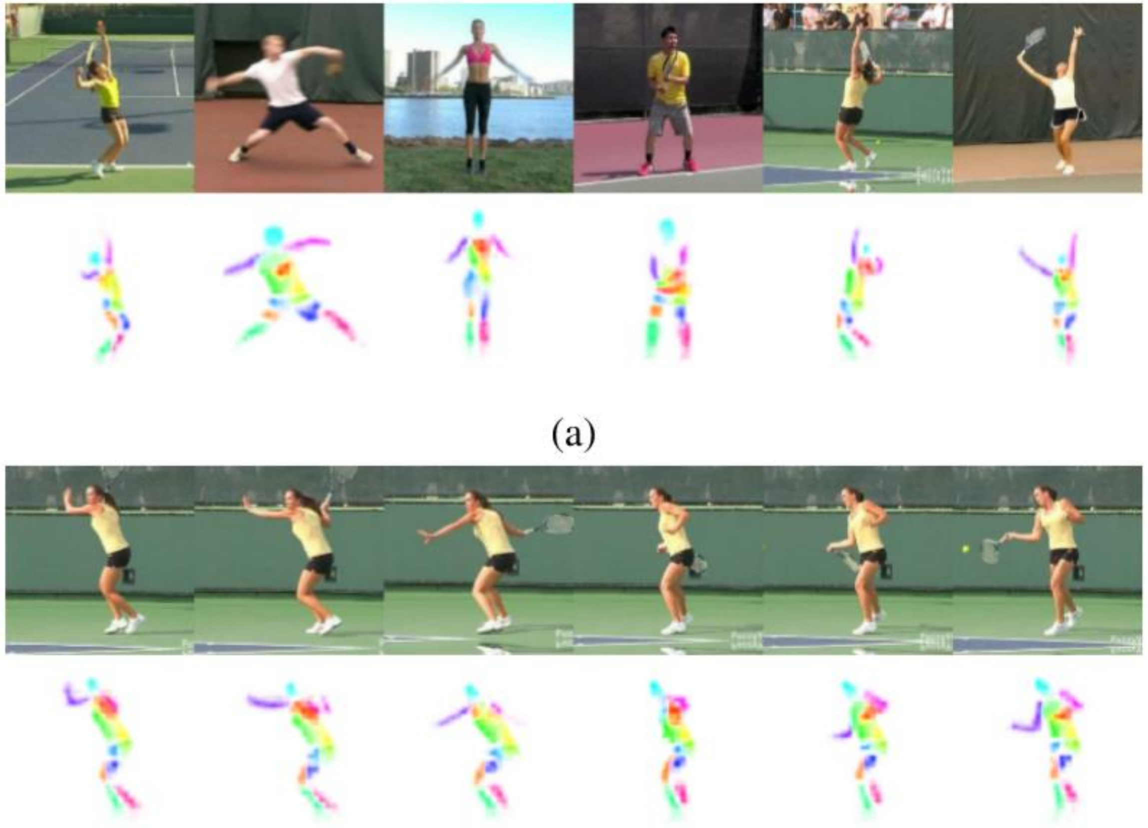
$$\mathcal{L}_{\text{rec}} = \left\| x - D\left(\left[\alpha_i(x \circ s), \sigma_i(a(x))\right]_{i=1,\dots}\right)\right\|_1$$

$$\mathcal{L}_{\text{equiv}} = \sum_i \lambda_\mu \|\mu[\sigma_i(x \circ s)] - \mu[\sigma_i(a(x)) \circ s]\|_2 + \lambda_\Sigma \|\Sigma[\sigma_i(x \circ s)] - \Sigma[\sigma_i(a(x)) \circ s]\|_1 \, ,$$

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{equiv}}$$

3 Experiments

Learned shape representation



(a) (b)

Figure 3: Learned shape representation on Penn Action. For visualization, 12 of 16 part activation maps are plotted in one image. (a) Different instances, showing intra-class consistency and (b) video sequence, showing consistency and smoothness under motion, although each frame is processed individually.

3 Experiments

Unsupervised Landmark Prediction

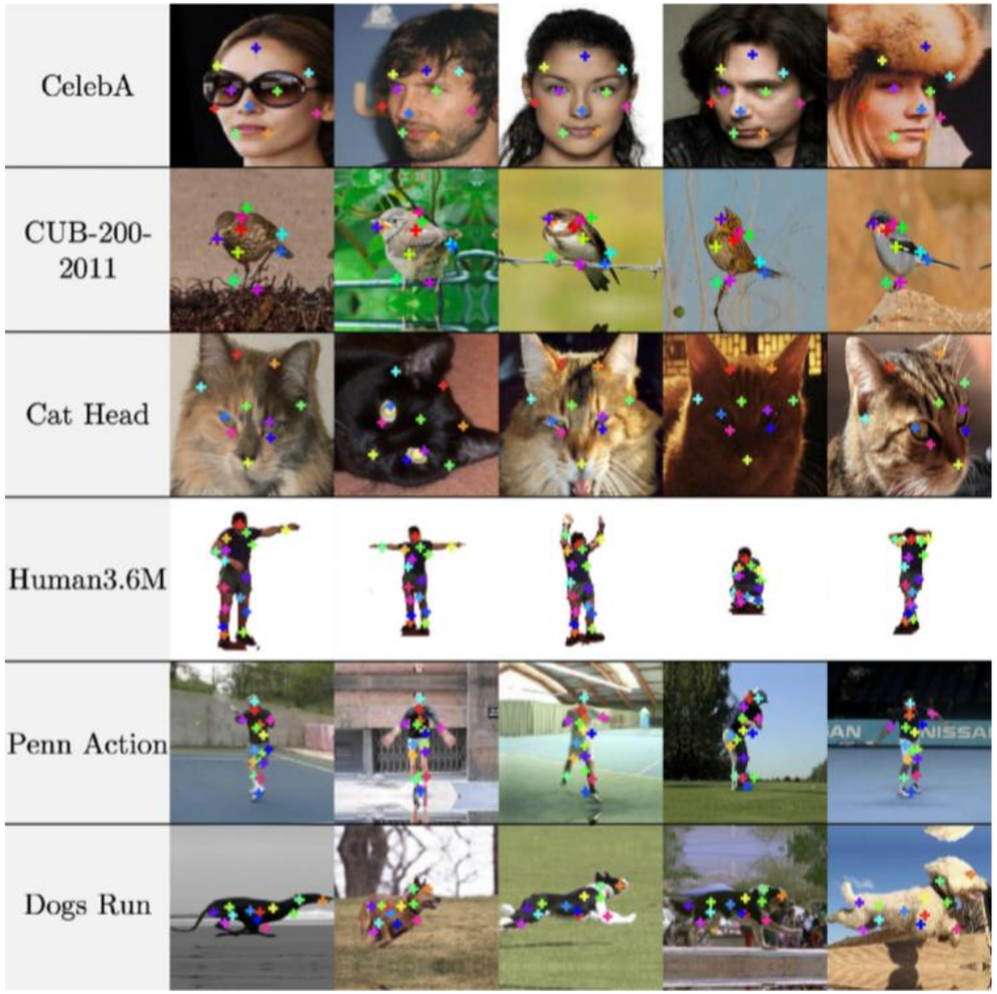


Table 2: Error of unsupervised methods for landmark prediction on the Cat Head, MAFL (subset of CelebA), and CUB-200-2011 testing sets. The error is in % of inter-ocular distance for Cat Head and MAFL and in % of edge length of the image for CUB-200-2011.

Dataset	Cat Head		MAFL	CUB
# Landmarks	10	20	10	10
Thewlis [47]	26.76	26.94	6.32	-
Jakab [22]	-	-	**3.19**	-
Zhang [60]	15.35	14.84	3.46	5.36
Ours	**9.88**	**9.30**	3.24	**3.91**

3 Experiments

Unsupervised Landmark Prediction

Table 3: Performance of landmark prediction on BBC Pose test set. As upper bound, we also report the performance of supervised methods. The metric is % of points within 6 pixels of groundtruth location.

BBC Pose		Accuracy
supervised	Charles [5]	79.9%
	Pfister [37]	88.0%
unsupervised	Jakab [22]	68.4%
	Ours	**74.5%**

Table 4: Comparing against supervised, semi-supervised and unsupervised methods for landmark prediction on the Human3.6M test set. The error is in % of the edge length of the image. All methods predict 16 landmarks.

Human3.6M		Error w.r.t. image size
supervised	Newell [33]	2.16
semi-supervised	Zhang [60]	4.14
unsupervised	Thewlis [47]	7.51
	Zhang [60]	4.91
	Ours	**2.79**

3 Experiments

Disentangling Shape and Appearance (Conditional Image Generation)

Table 5: Mean average precision (mAP) and rank-n accuracy for person re-identification on synthesized images after performing shape/appearance swap. Input images from Deep Fashion test set. Note [13] is supervised w.r.t. shape.

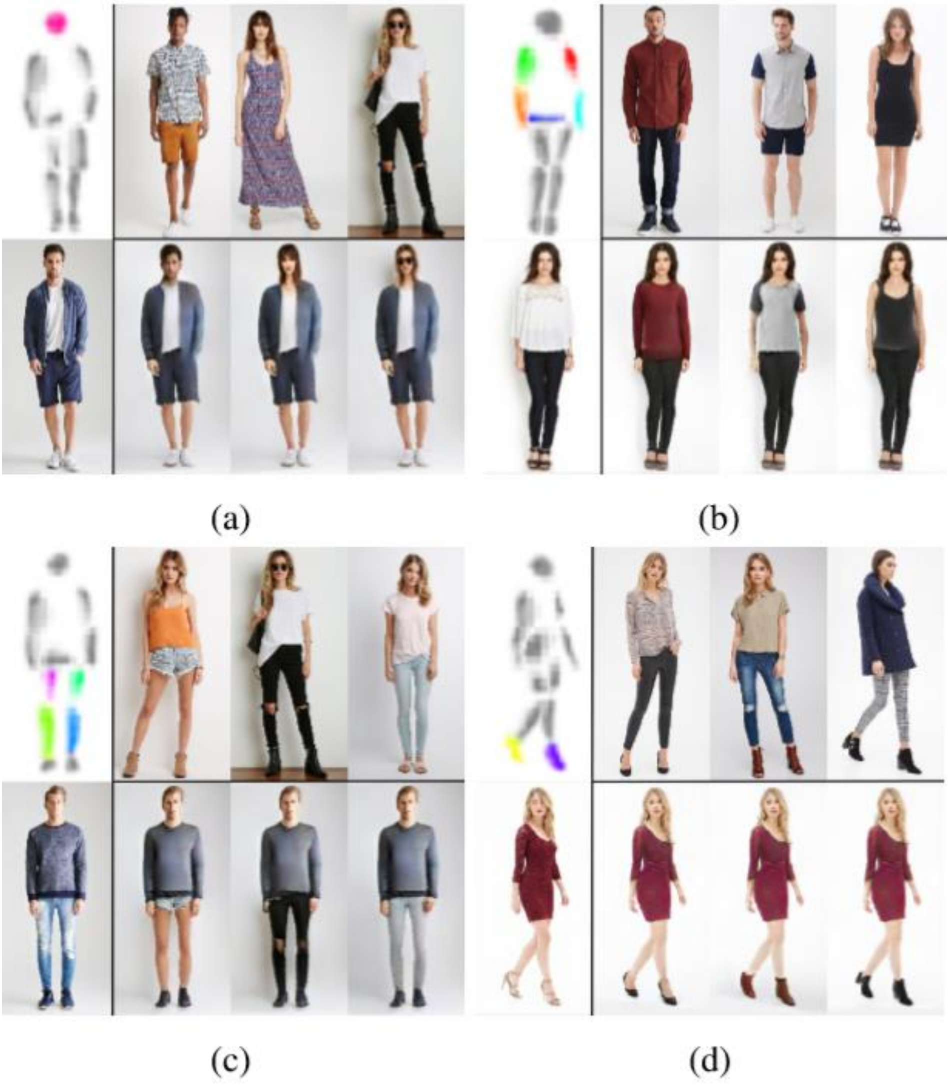
	mAP	rank-1	rank-5	rank-10
VU-Net [13]	88.7%	87.5%	98.7%	99.5%
Ours	90.3%	89.4%	98.2%	99.2%

Table 6: Percentage of Correct Keypoints (PCK) for pose estimation on shape/appearance swapped generations. α is pixel distance divided by image diagonal. Note that [13] serves as upper bound, as it uses the groundtruth shape estimates.

α	2.5%	5%	7.5%	10%
VU-Net [13]	95.2%	98.4%	98.9%	99.1%
Ours	85.6%	94.2%	96.5%	97.4%

3 Experiments

Disentangling Shape and Appearance (Part Appearance Transfer)



3 Experiments

Disentangling Shape and Appearance (Video-to-Video Translation)

https://compvis.github.io/unsupervised-disentangling/