Hierarchical Part-based Disentanglement of Pose and Appearance

M.Sc. Thesis presentation by ***Farnoosh Javadi***

Supervisors: Jim Little, Helge Rhodin

THE UNIVERSITY OF BRITISH COLUMBIA

December 2020

Overview

Motivation

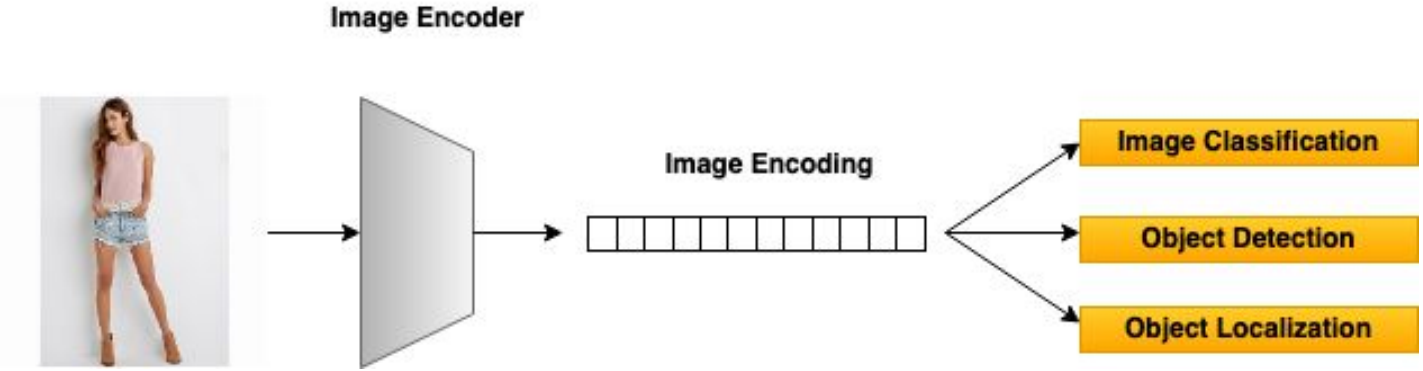
Intuition
Problem formulation
PD (Part-based Disentanglement)
HPD (Hierarchical Part-based Disentanglement)

Summary
Future work

Introduction

Motivation

- Learning Image Representations
  - a crucial task in Computer Vision

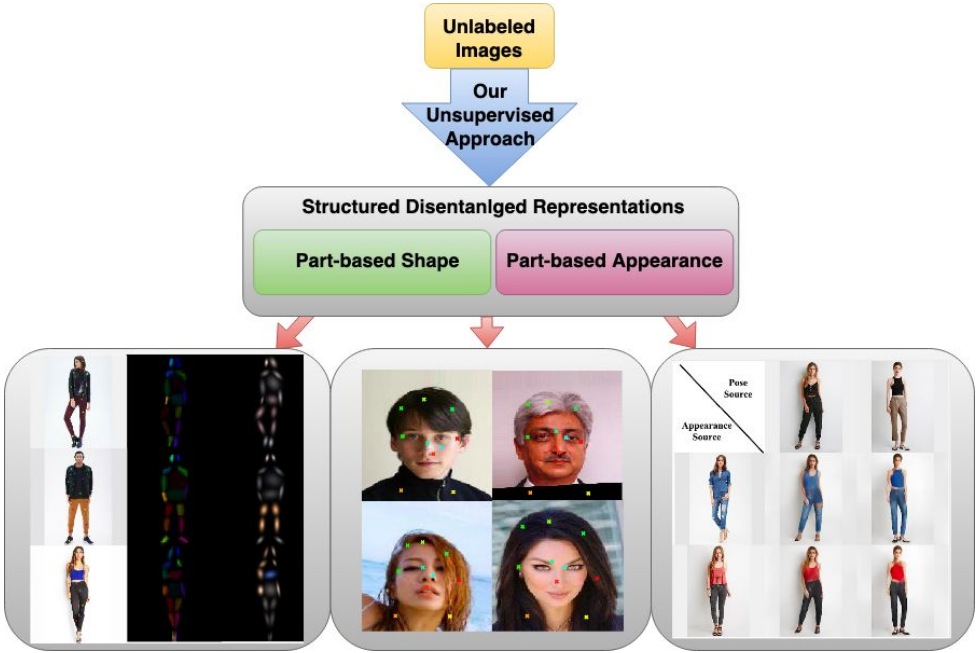


- **Problem**: Entangled Representations lack interpretability
- **Solution:** Disentangled Representations
  - Factorized
  - Each factor represents an independent characteristic of objects.
  - More interpretable
  - Enable Novel Image Synthesis

Our Project

- Our representations encode:
  - Object keypoints (parts)
  - Disentangled appearance and shape of parts
  - Parts structure in the shape of a binary-tree

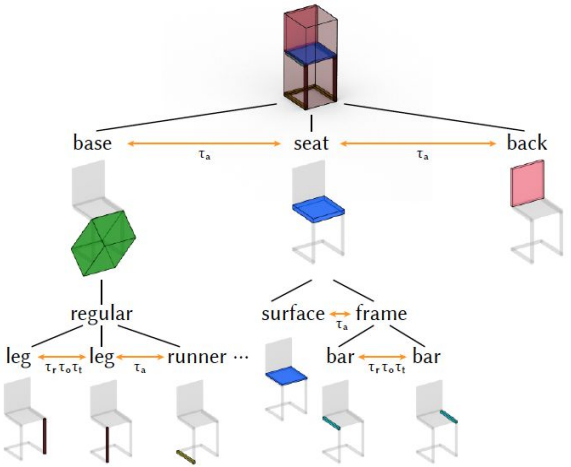
- Our provided latent space can be used for:
  - Landmark detection
  - Selective image modification
  - Local appearance or pose transfer



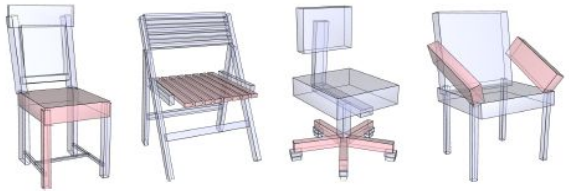
Related Work

Supervised Approaches

- Need annotations in terms of:
  - Pose annotations
  - Keypoints
  - Segmentation of shape into parts
  - Parts labels
  - Hierarchy of parts
- Cons:
  - Tons of annotations
  - Hard to get annotations
  - Not applicable for domains with no labels
- Our method:
  - Unsupervised
  - Needs no annotation



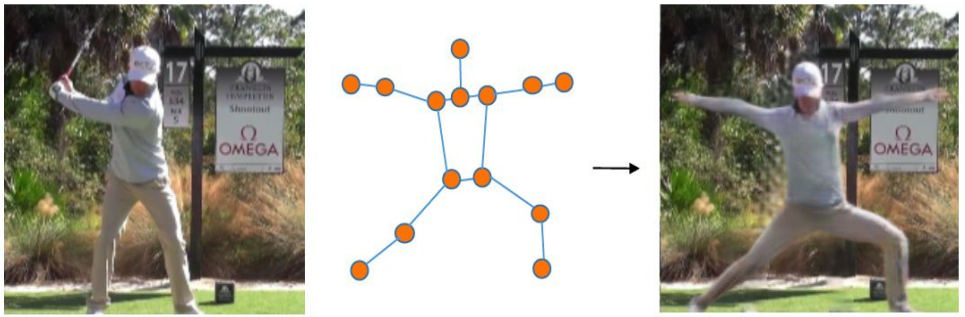
Paschalidou et al.



Li et al.

Unsupervised Approaches

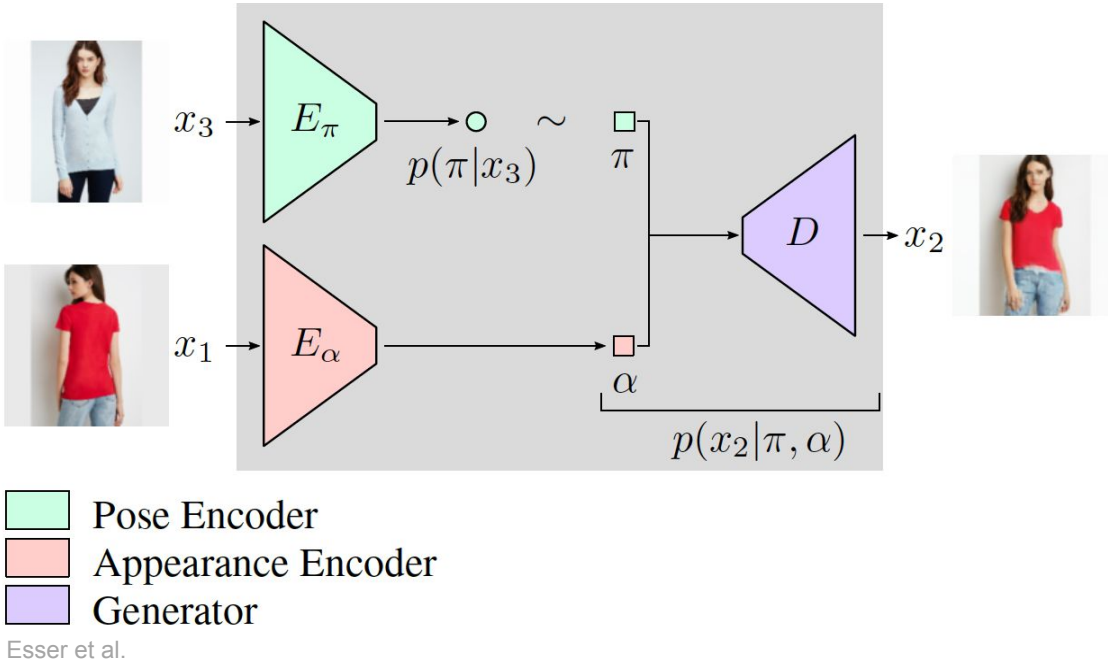
- Condition generative models on pose or keypoints information
  - Extracted from a pre-trained detector
- Cons:
  - Only applicable when a pose detector exists
  - Not generalizable to all objects
- Our method :
  - No prior knowledge about object shape
  - Applicable for any arbitrary domain.



Balakrishnan et al.

Approaches Trained on Multiple Images

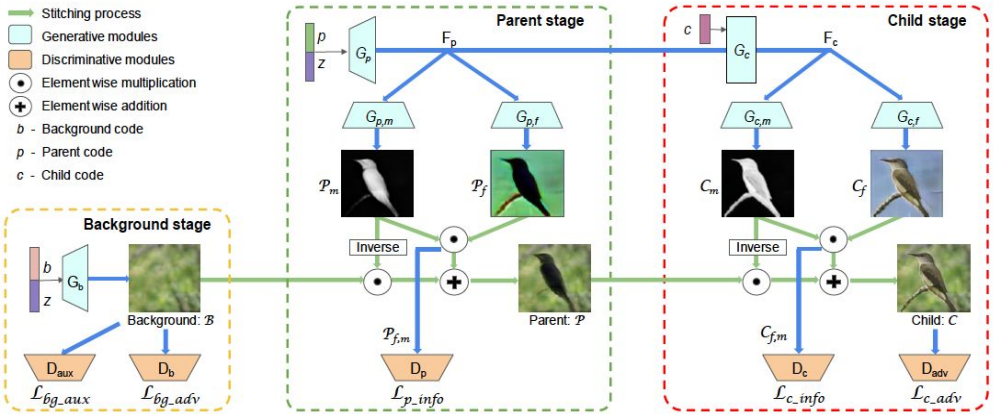
- Some methods need:
  - Multiple frames from videos
  - Pairs of images
    - Varying in one factor
    - Same in the rest
- Cons:
  - Such datasets are hard to obtain
- Our method:
  - Trained on single images
  - Applicable for videos



Esser et al.

GAN-based methods

- Train with adversarial loss
- Cons:
  - Hard to optimize
  - Can't encode existing images
- Our method:
  - Auto-encoder
  - Robust training with reconstruction loss



Singh et al.

Method

Intuition

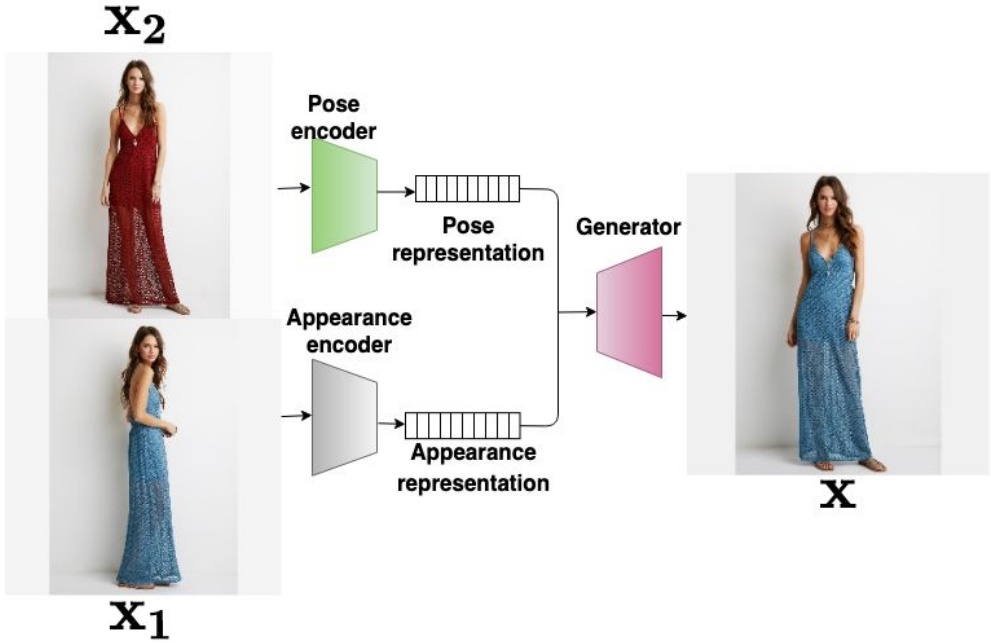
- Let us assume we have a triplet of images (x, x1, x2)
  - x and x1 share the same appearance
  - x and x2 share the same pose

  - An Image can be generated by
    - Its pose and appearance :

$$x = D(\Phi^{app}(x), \Phi^{pose}(x)).$$

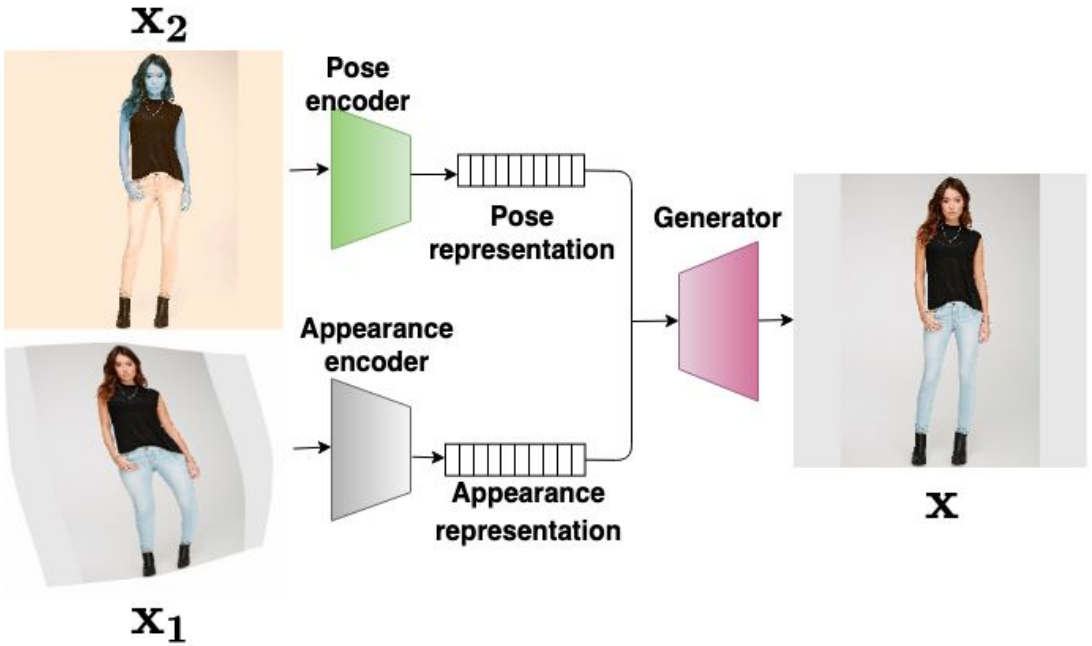
- The above formula can be re-written :

$$x = D(\Phi^{app}(x_1), \Phi^{pose}(x_2)).$$



Our Method

- We **create** triplets from single images
  - x1 → Spatially transformed image
  - x2 → Appearance transformed image



Transformations

- Shift in HSV space
- Mix with a base color in RGB space

- 7 pre-defined TPS
- Linearly combine them
- Rotate up to 60 degrees

Problem Formulation

- Factorize image into its forming parts

$$\Phi(x) = (\Phi_1(x), \Phi_2(x), \ldots, \Phi_k(x))$$

- Part representation should be a combination of its appearance and pose

$$\Phi_i(x) = (\Phi_i^{app}(x), \Phi_i^{pose}(x))$$

- Pose and Appearance should be invariant to changes in appearance and pose

$$\Phi_i(x) = (\Phi_i^{app}(T_s(x)), \Phi_i^{pose}(T_a(x)))$$

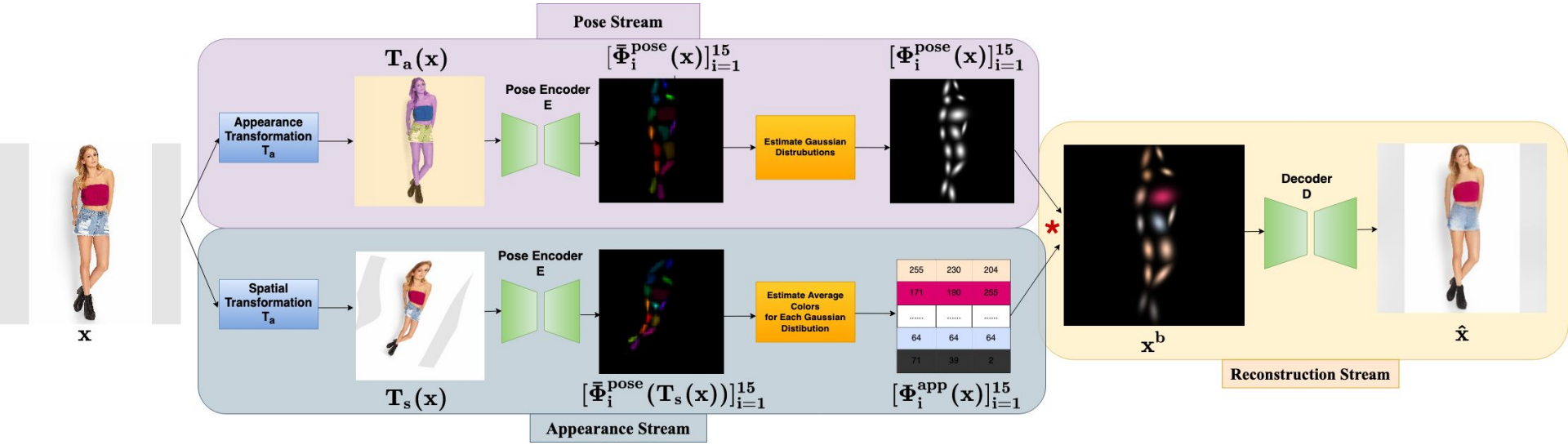
- Reconstruction is an assembly of parts shape and appearance

$$\Phi(x) = [(\Phi_i^{app}(T_s(x)), \Phi_i^{pose}(T_a(x)))]_{i=1}^{k}$$

PD

(Part-based Disentanglement)

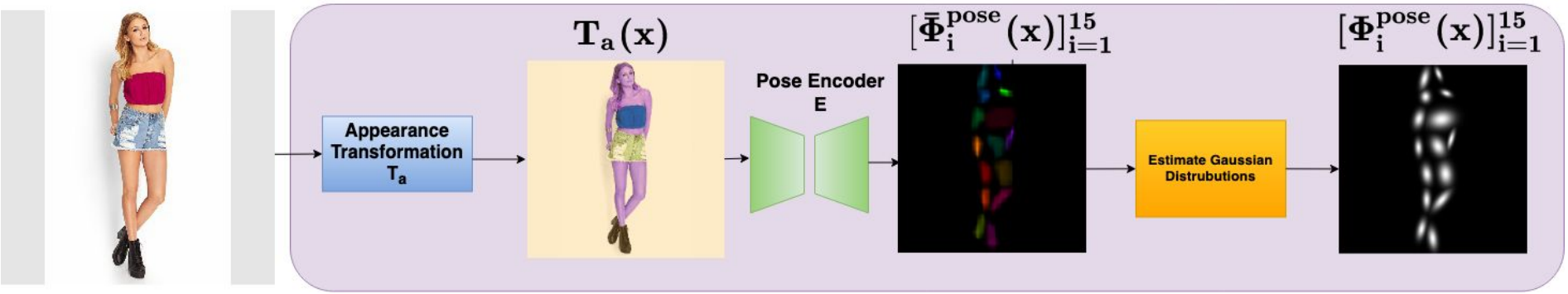
PD


Pose Stream
$\mathbf{T_a(x)}$
$[\bar{\Phi}_i^{pose}(x)]_{i=1}^{15}$
$[\Phi_i^{pose}(x)]_{i=1}^{15}$
Appearance Transformation T_a
Pose Encoder E
Estimate Gaussian Distrubutions
Spatial Transformation T_a
Pose Encoder E
Estimate Average Colors for Each Gaussian Distibution
255 230 204
171 190 255
64 64 64
71 39 2
$\mathbf{T_s(x)}$
$[\bar{\Phi}_i^{pose}(T_s(x))]_{i=1}^{15}$
$[\Phi_i^{app}(x)]_{i=1}^{15}$
Appearance Stream
Decoder D
$\mathbf{x}$
$\mathbf{x^b}$
$\hat{\mathbf{x}}$
Reconstruction Stream

Pose Stream

- **Goal**: Predict object parts in terms of 2D Gaussian distributions.

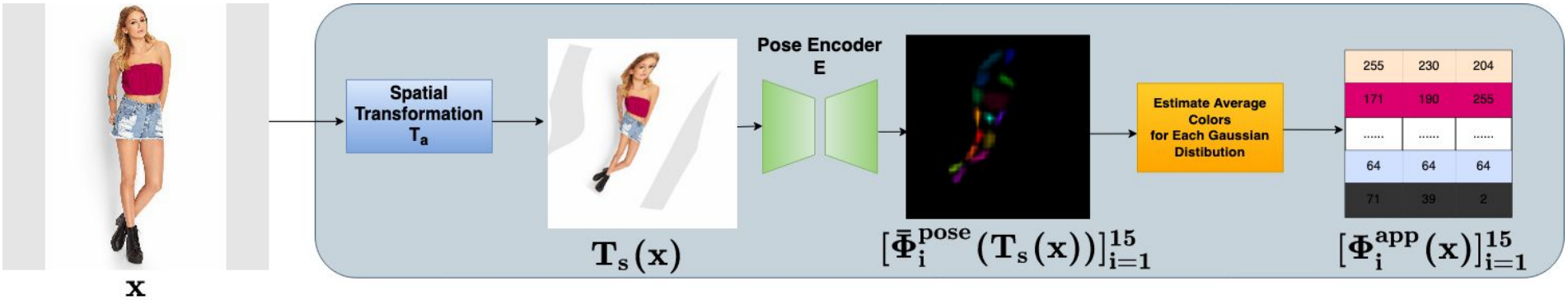
$$\Phi^{pose}(x) = [\Phi_i^{pose}(x)]_i$$



Appearance Stream

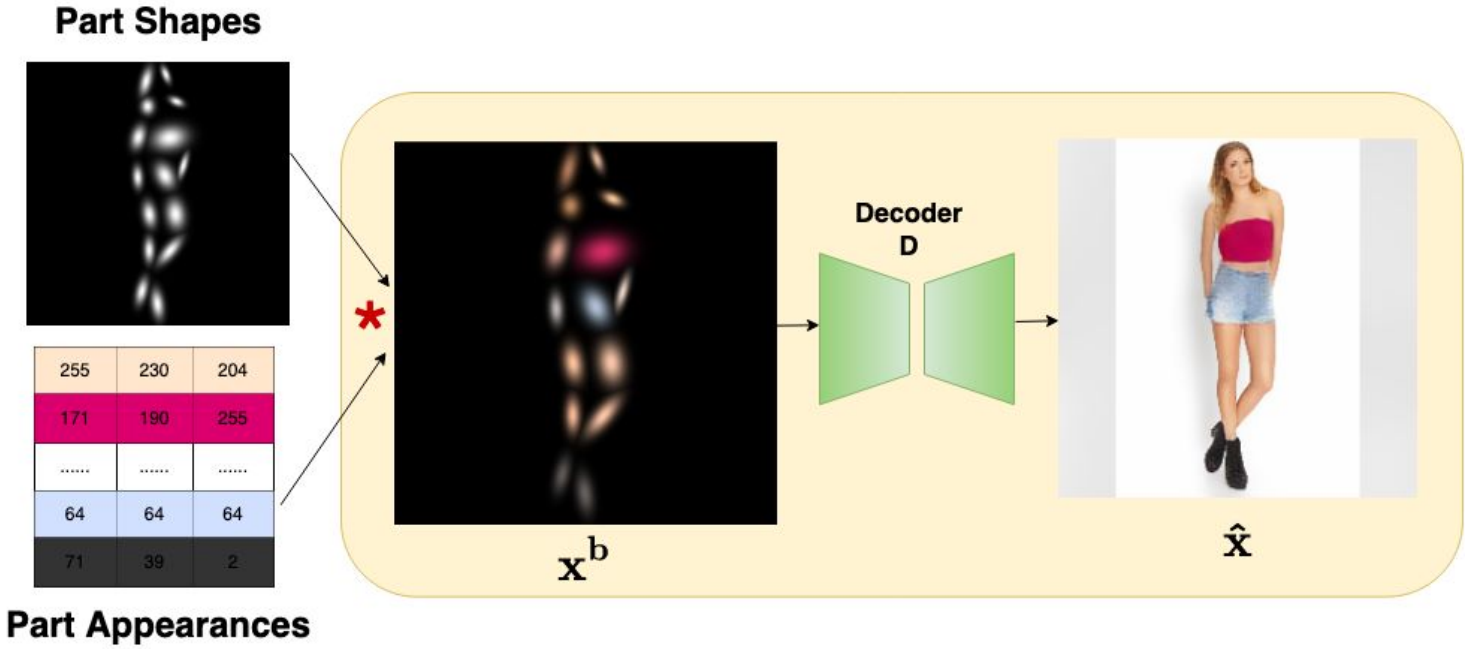
- **Goal**: Predict appearance vectors for each part

$$\Phi^{app}(x) = [\Phi_i^{app}(x)]_{i=1}^{k}$$



Reconstruction Stream

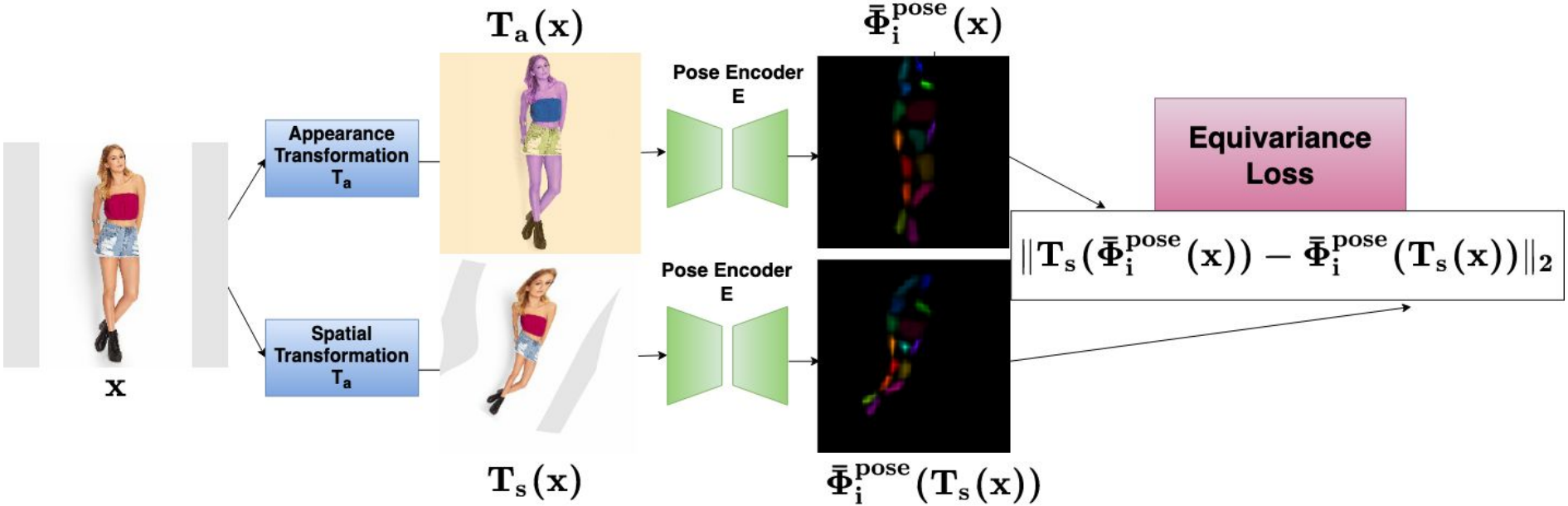
- **Goal**: Reconstruct image from the set of parts and their appearances.



Training Objective

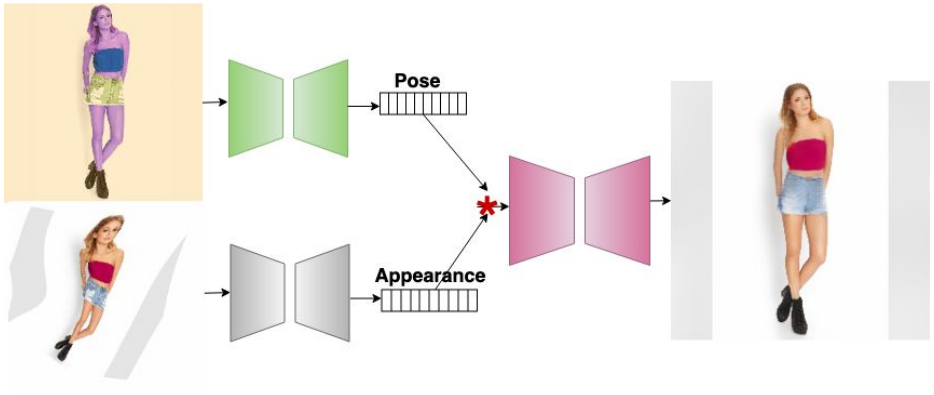
- $l_{\text{rec}} = \|x - \hat{x}\|_2 + l_{\text{perc}}(x, \hat{x})$

Equivariance Constraint

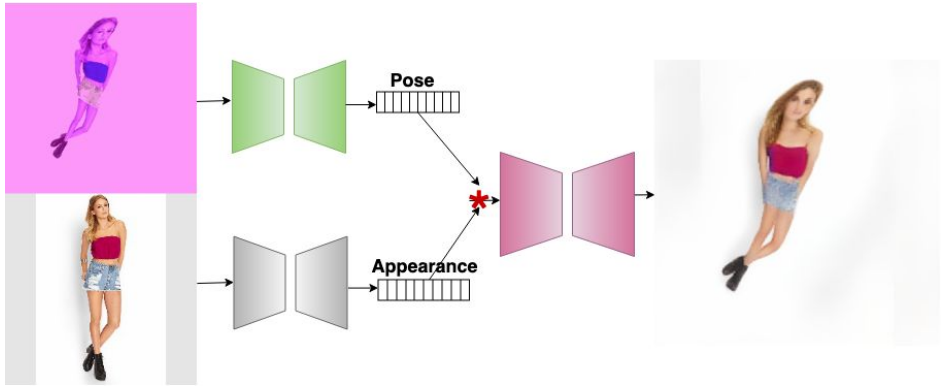


Swapping Technique

- Half of the times, we randomly swap role of the original image and the spatially transformed image.

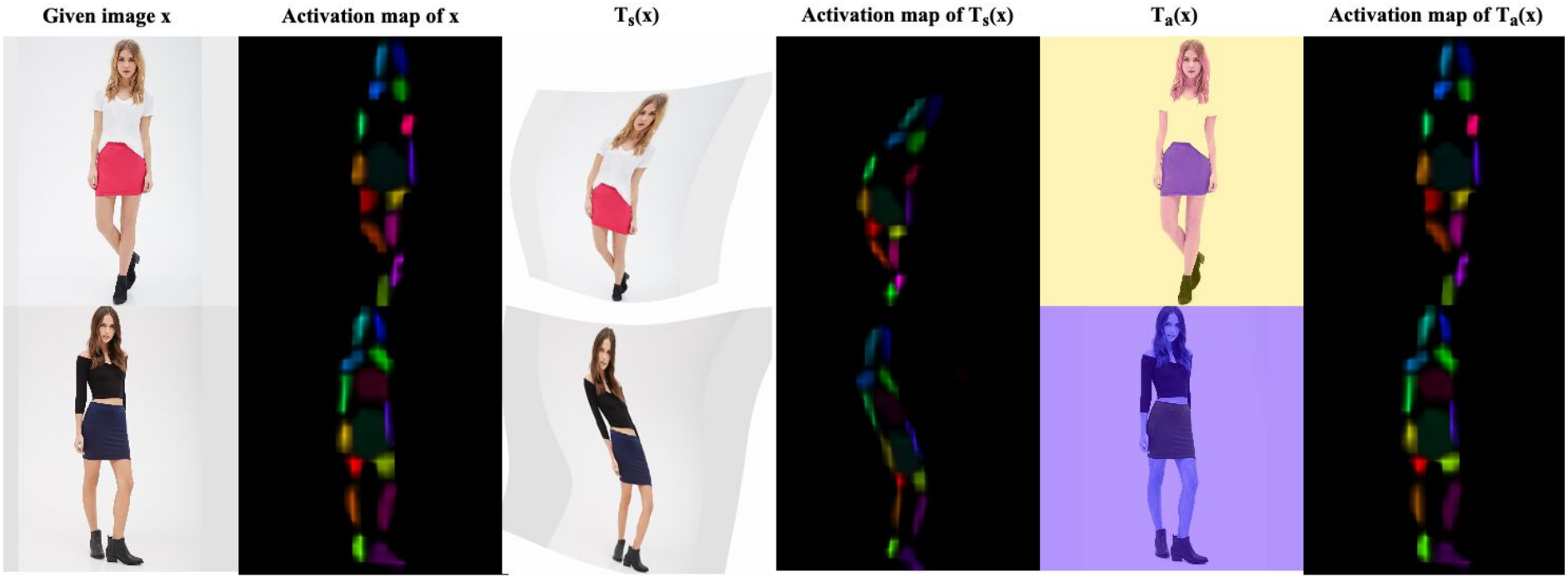


Half of the times



Half of the times

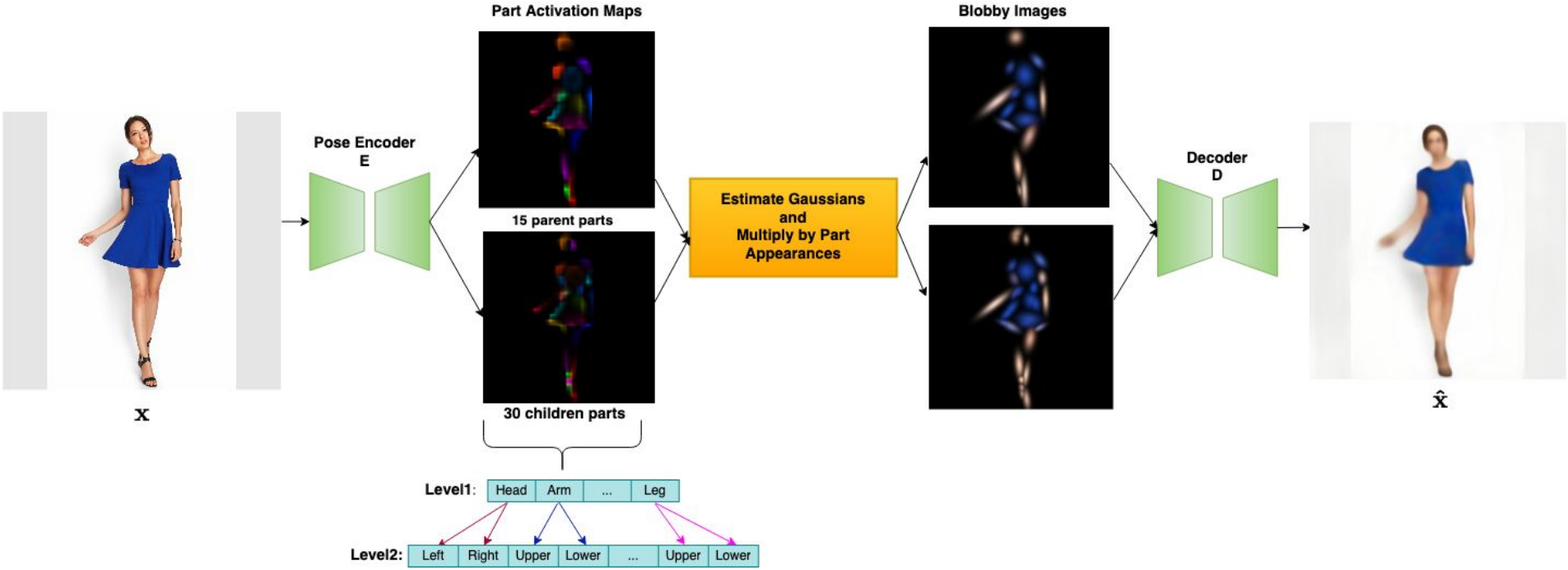
Pose Invariance and Equivariance



HPD

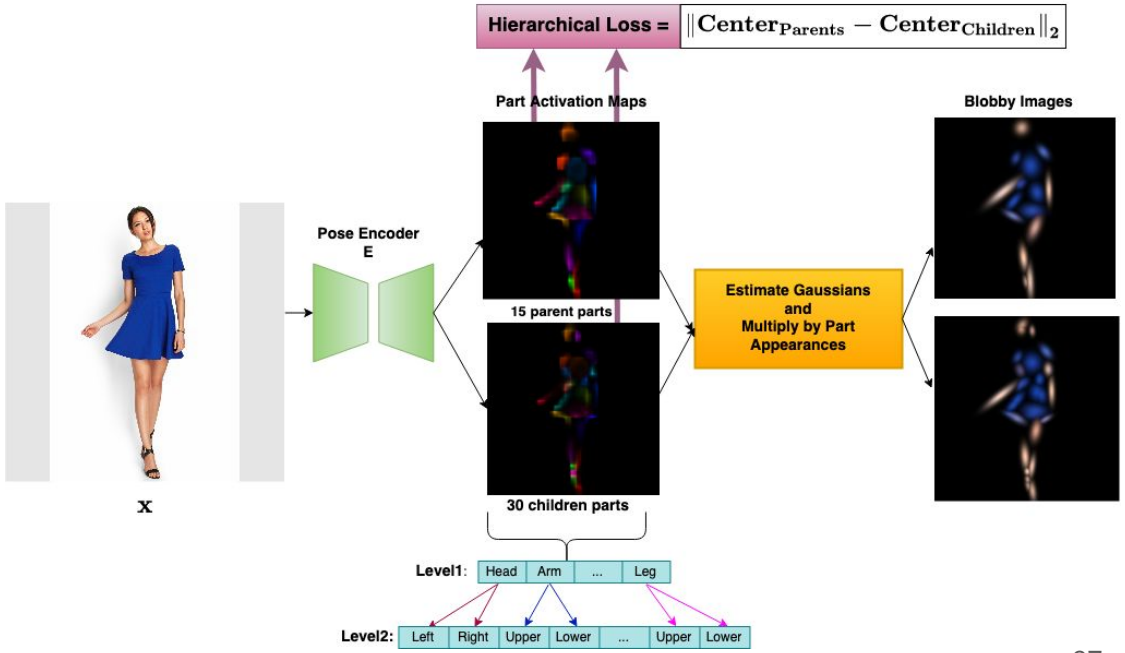
(Hierarchical Part-based Disentanglement)

HPD


Part Activation Maps
Blobby Images
Pose Encoder
E
15 parent parts
Estimate Gaussians
and
Multiply by Part
Appearances
Decoder
D
30 children parts
x
x̂
Level1:
Head
Arm
...
Leg
Level2:
Left
Right
Upper
Lower
...
Upper
Lower

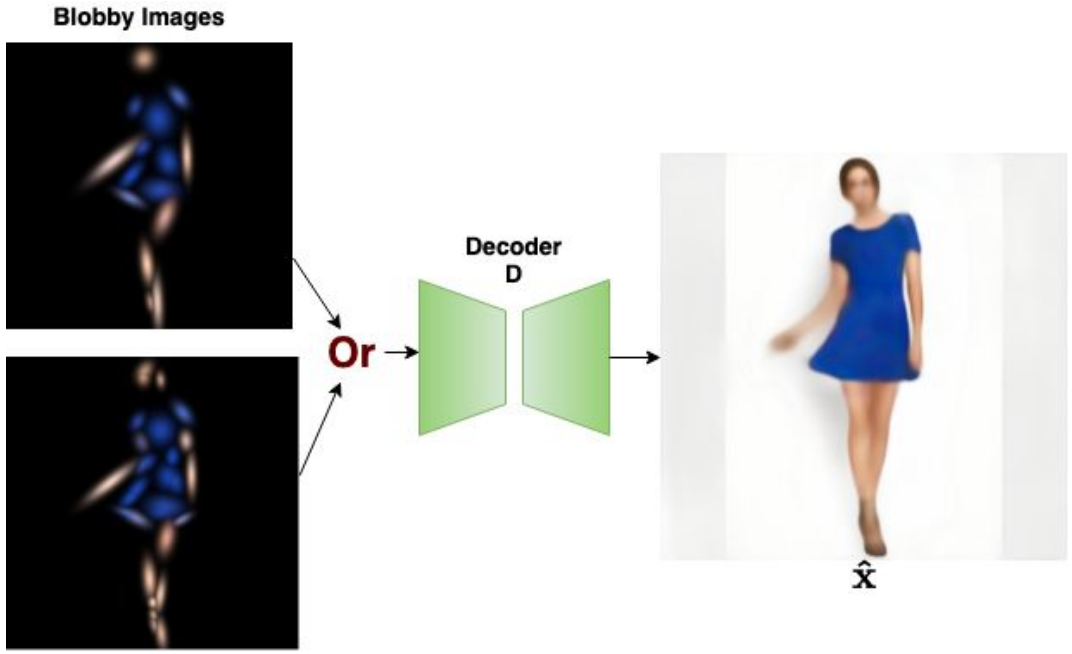
Pose and Appearance Stream

- Pose stream
  - **Goal**: Learn structured parts
- Appearance stream
  - **Goal**: Learn the color of each part



Reconstruction Stream

- We randomly pick one set of parts either the parents or the children for reconstruction.



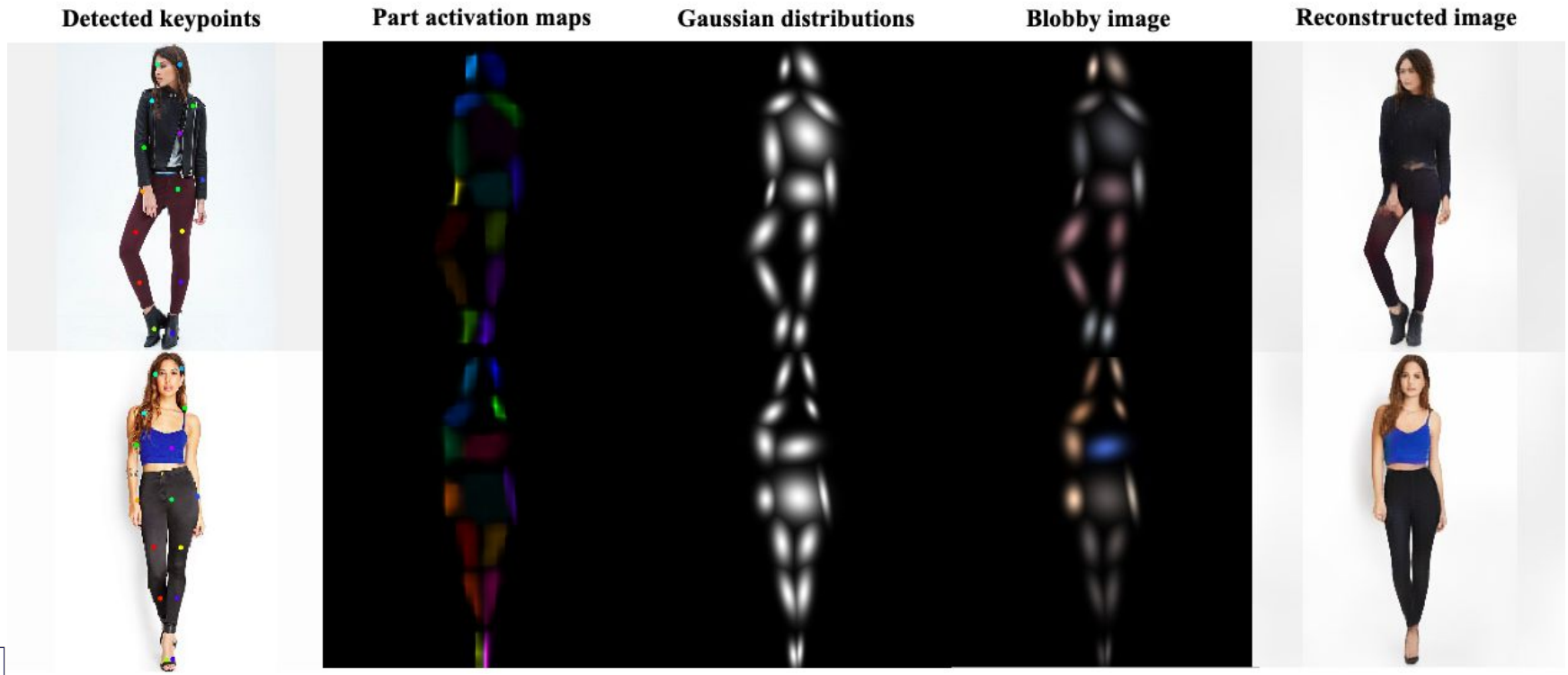
Transformations

- Appearance Transformations
  - Shift in HSV space
  - Mix with a base color in RGB space

- Spatial Transformations
  - 7 pre-defined TPS
  - Linearly combine them
  - Rotate up to 60 degrees

Results

Part Detection

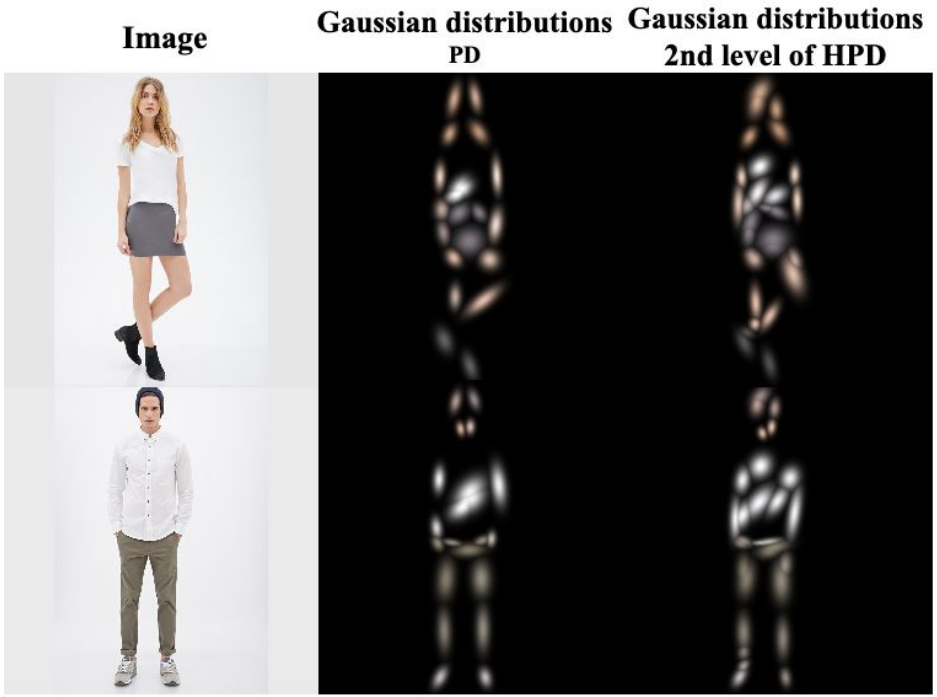


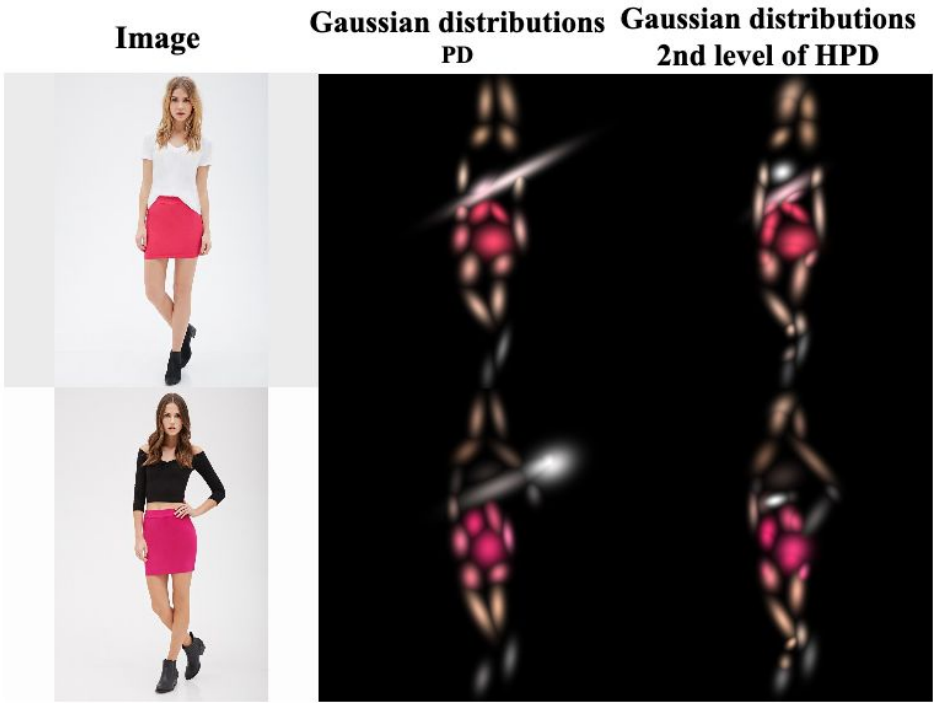
Global Pose and Appearance Transfer

Local Appearance Transfer

Local Pose Transfer

PD vs HPD - Qualitative



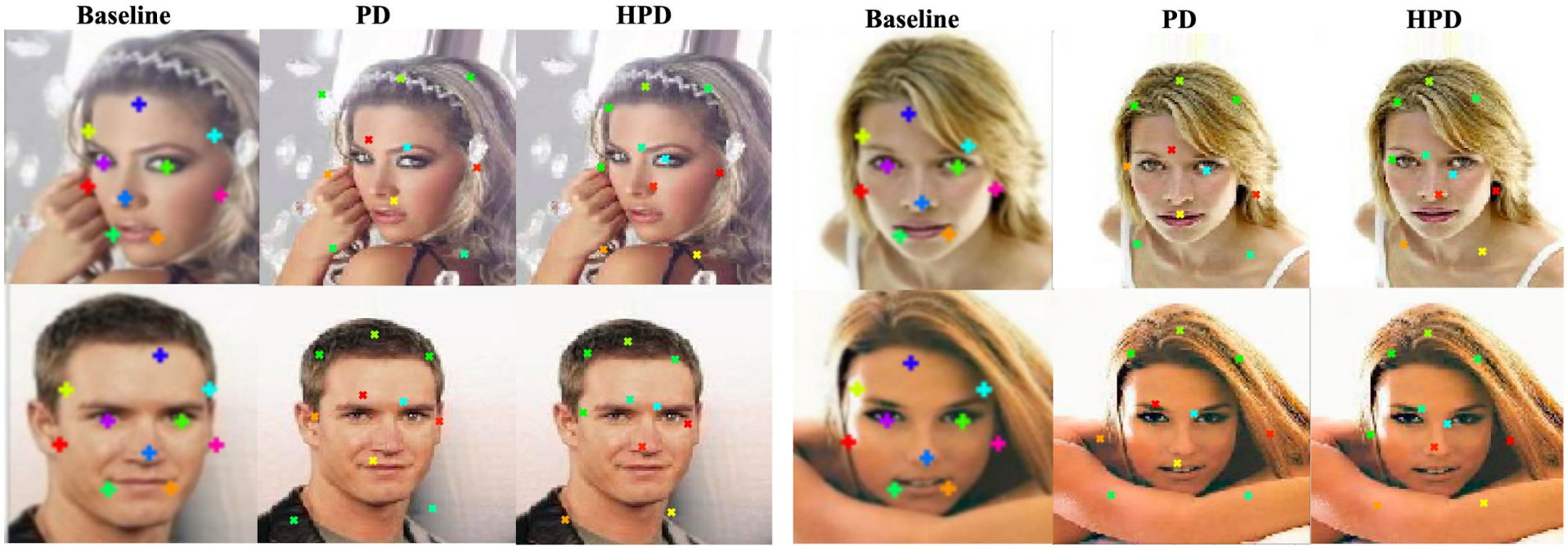


PD vs HPD - Quantitative

- HPD predicts more meaningful and detailed parts that lead to better reconstruction
  - In terms of pixel-wise error
  - On the DeepFashion dataset

Method	Pixel-wise Error
PD	0.2921
HPD	**0.2202**

Unsupervised Landmark Detection



Baseline (Lorenz et al.)

Regressed Keypoints

Ablation Study

- We compared 5 variations of the model each differing in one module.
  - In terms of Landmark Detection Error
  - On CelebA dataset

- HPD yielded the best results

Method	Model's Parameters	Landmark Detection Error
Baseline (Lorenz et al.)	74,171,543	7.54 (3.24)
Baseline + Swapping	74,171,543	7.12
Baseline + RGB colors	56,903,565	6.25
PD	56,903,565	5.87
HPD	56,904,850	**5.79**

Conclusion

Summary

- An approach for detecting object parts and their disentangled appearance and pose in a **hierarchical** manner.
  - Unsupervised
  - Needs no prior knowledge about the object shape
  - Trained on single images
- Contributions:
  - Swapping Technique → Frees us from further tuning training objective
  - Simple appearance encoding method → Saves millions of parameters
  - Hierarchical extension → Detects more meaningful parts
- Evaluation
  - Part detection
  - Pose and appearance transfer
  - Landmark detection

Limitations and Future Work

- Enhance visual quality of images
  - Adding adversarial loss
- Test other datasets
  - Cats
  - Birds
  - Videos
- Do more experiments on the hierarchy
  - Assess the impact of depth
  - Increase number of landmarks
  - Enforce an unbalanced tree structure

Thanks for your Attention!