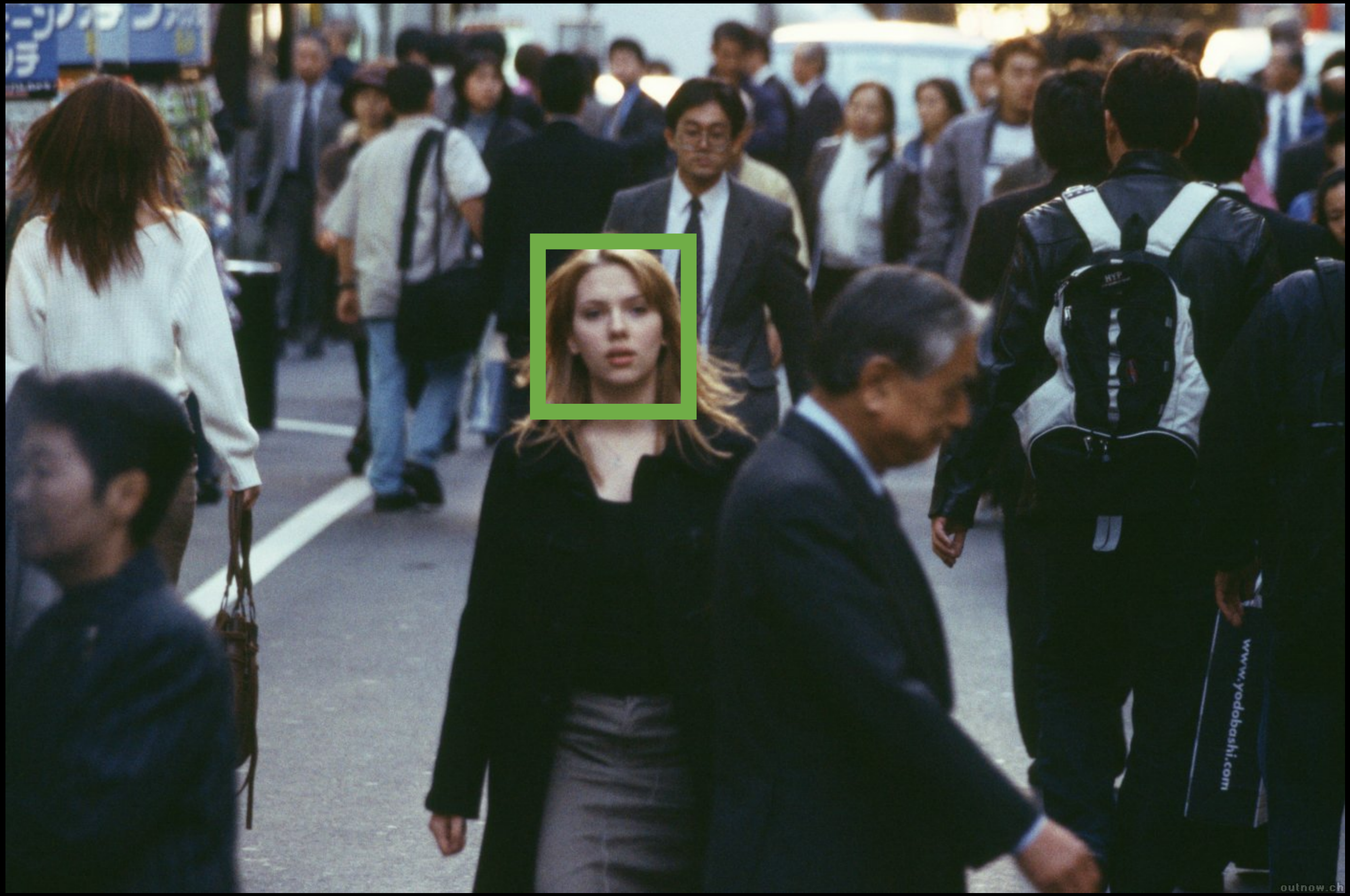
Deep Learning Human-centric Representation in the Wild

Ziwei Liu

The Chinese University of Hong Kong
University of California, Berkeley

Human-centric Analysis

Human-centric Analysis (I)



Face Understanding

Human-centric Analysis (II)

Fashion Understanding

Human-centric Analysis (III)

Scene Understanding

Human-centric Analysis (IV)

Motion Understanding

Part I: Deep Face Understanding

“Deep Learning Face Attributes in the Wild”, *ICCV 2015*

Face Attributes Recognition

- Problem

Arched Eyebrows?

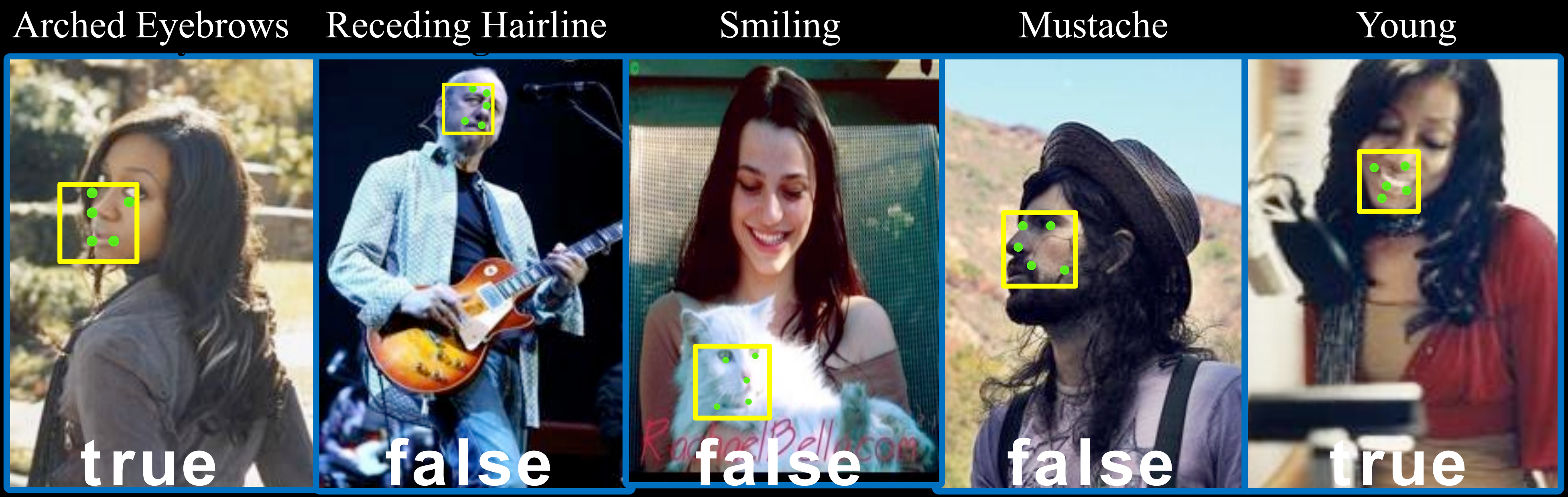
Big Eyes?

Receding Hairline?

Mustache?

Face Attributes Recognition

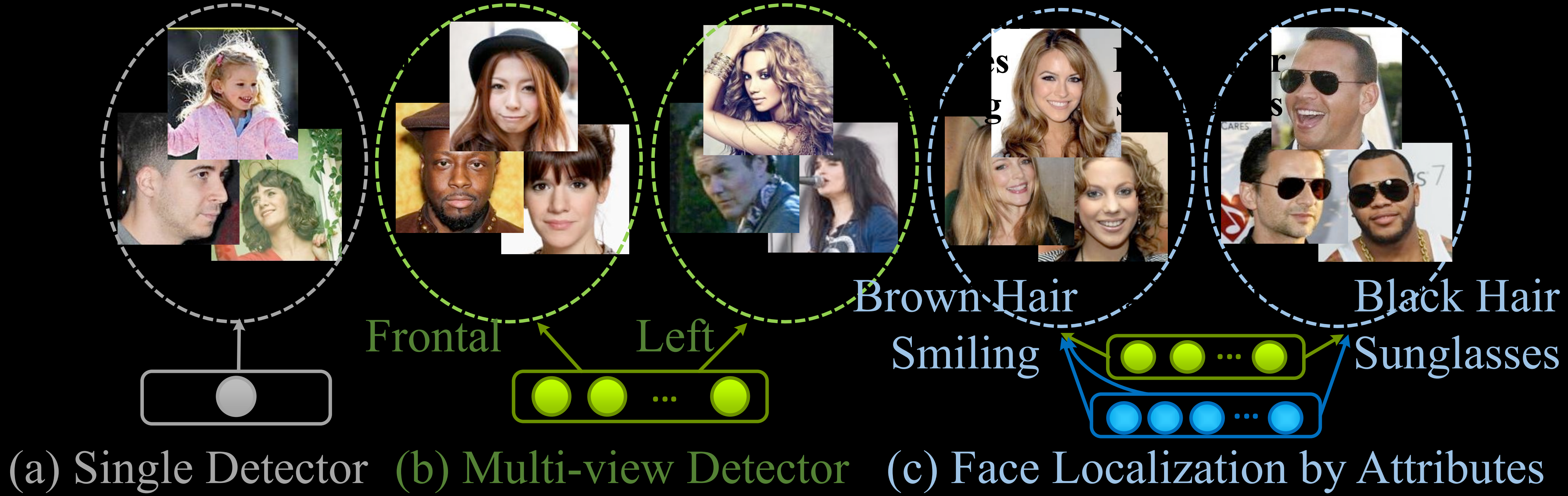
- Challenges



HOG (landmarks) + SVM

Face Attributes Recognition

- Motivation



CelebA

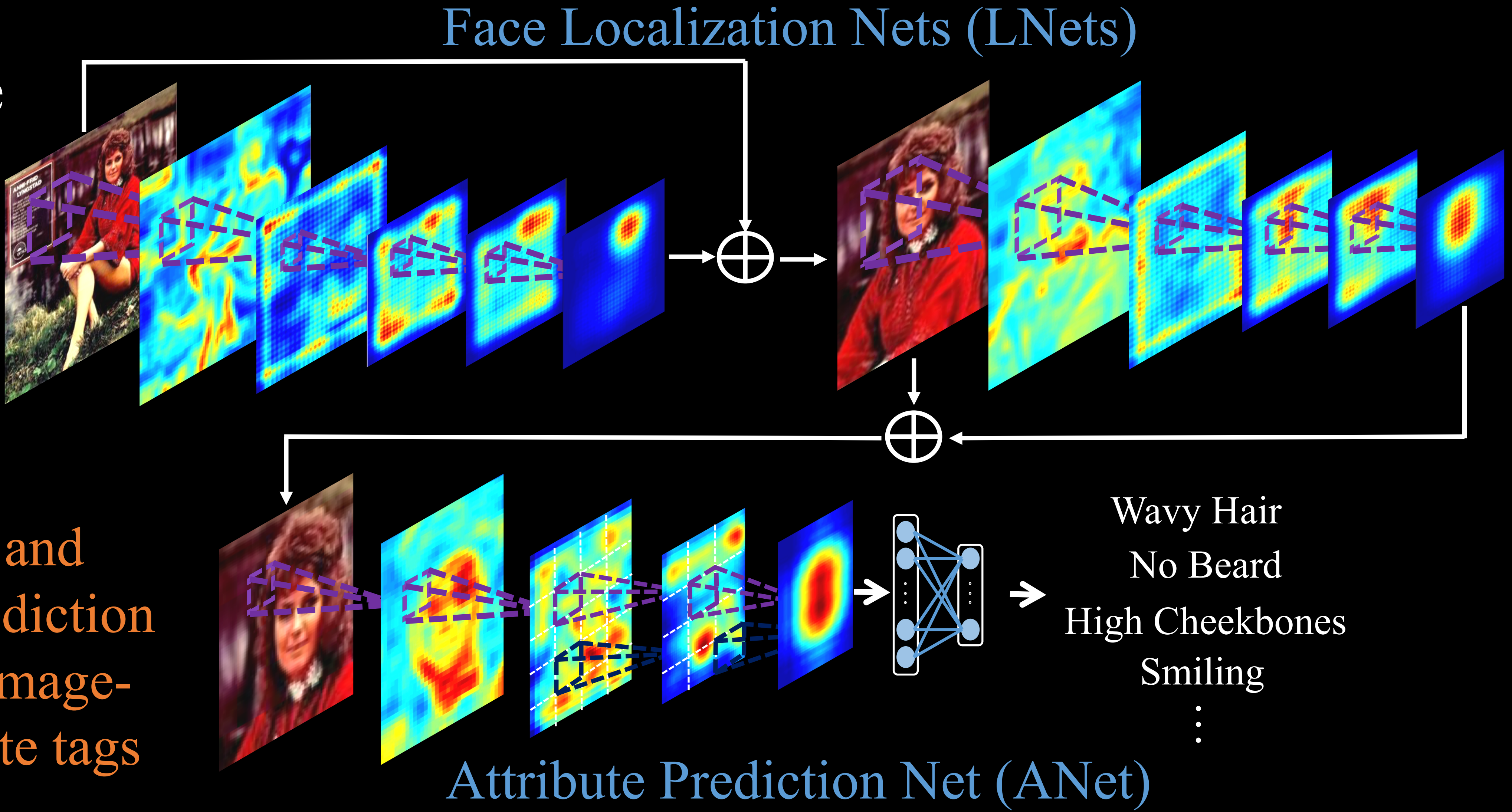
- 200,000 images

- 40 attributes
- 10,000 identities

- 1 bounding box
- 5 landmarks

Face Attributes Recognition

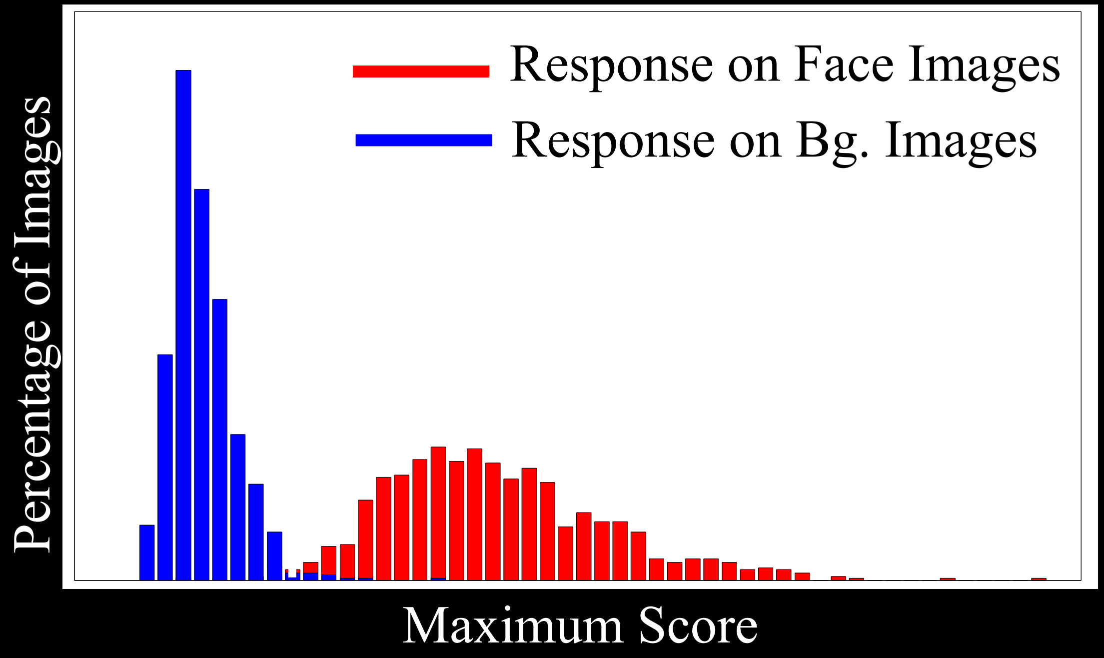
- Pipeline

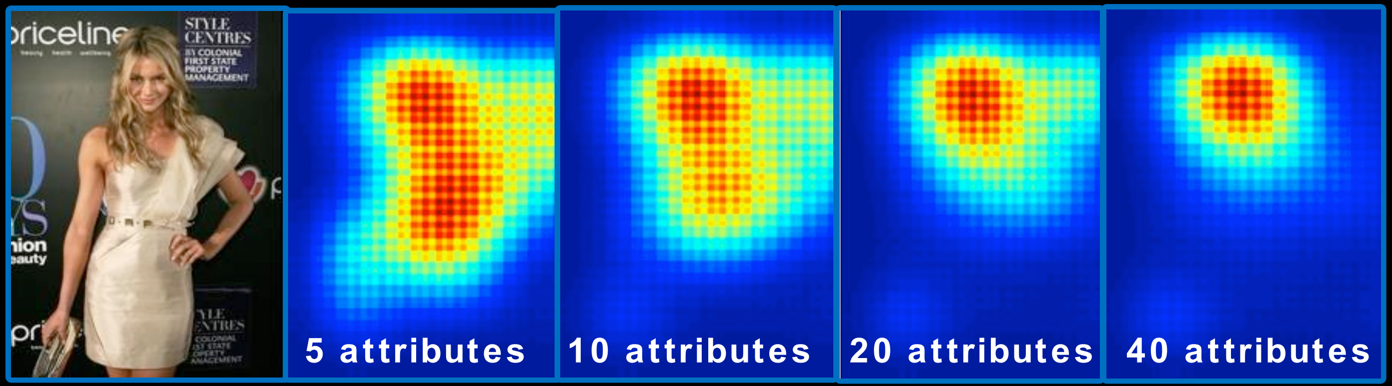


joint face localization and attribute prediction using only image-level attribute tags

Deep Face Representation

- Rich attributes tags enable accurate face localization





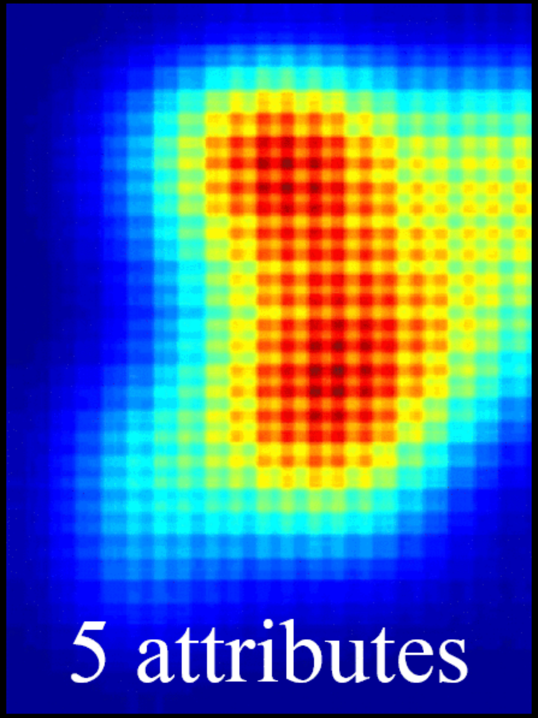
Response map with different numbers of attributes

Deep Face Representation

- Rich attributes tags enable accurate face localization

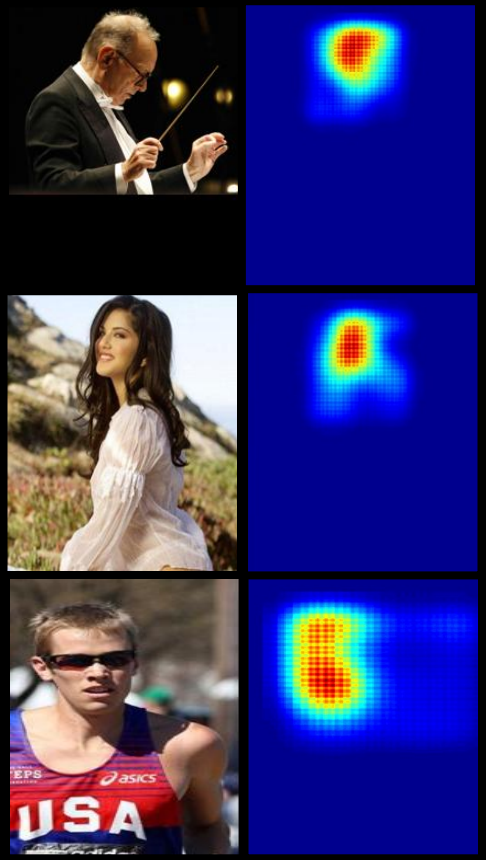
Original Image

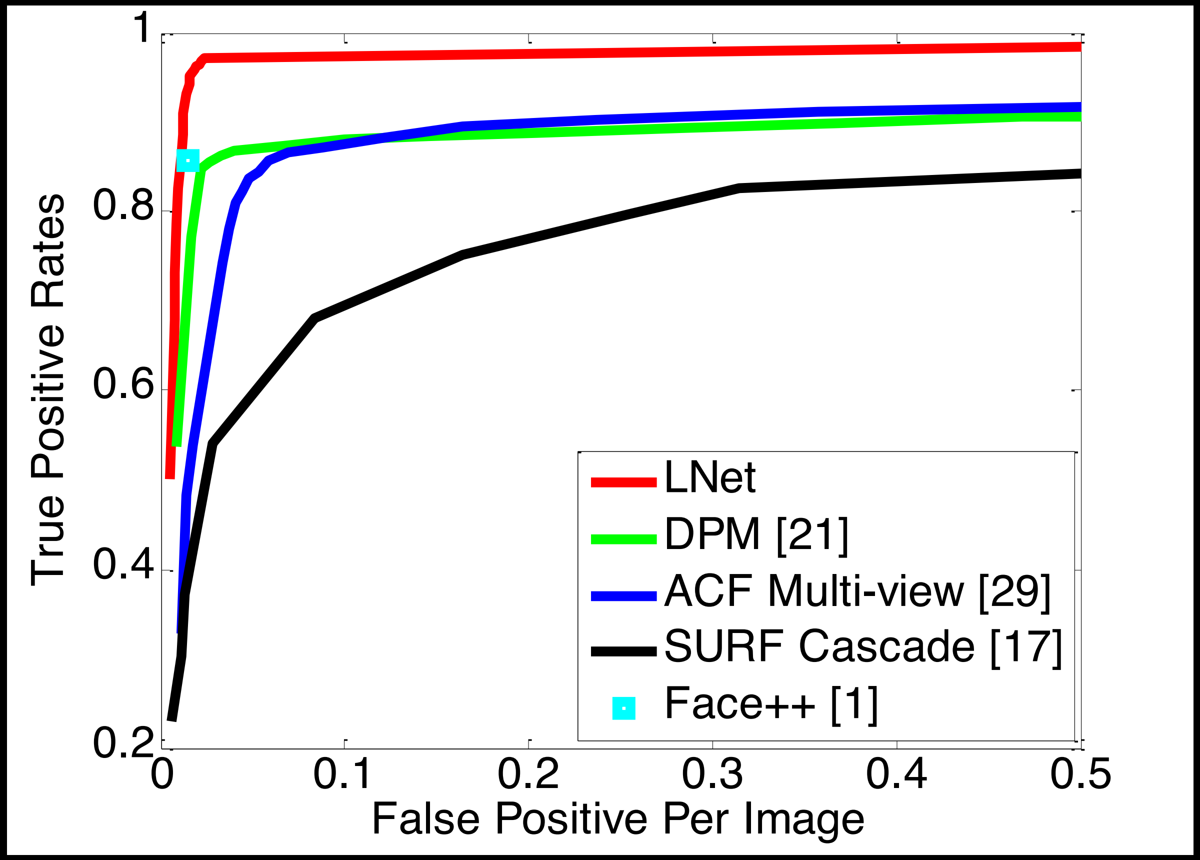
Response Map



Deep Face Representation

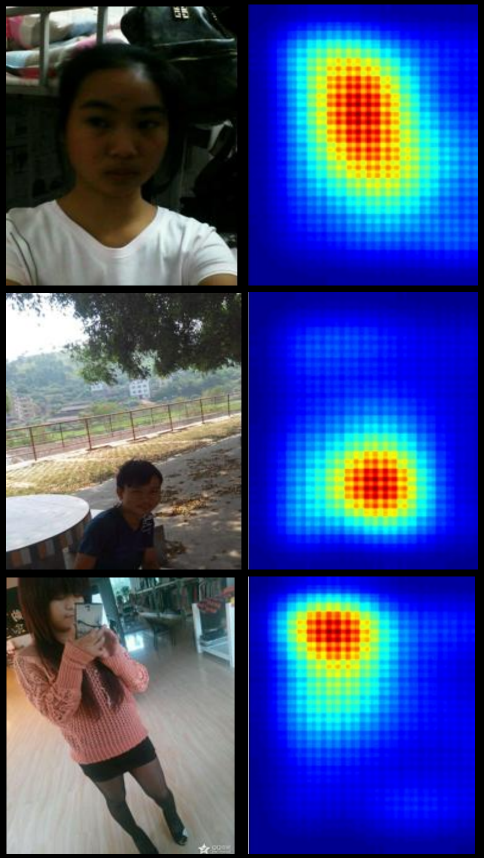
- Face localization performance on CelebA

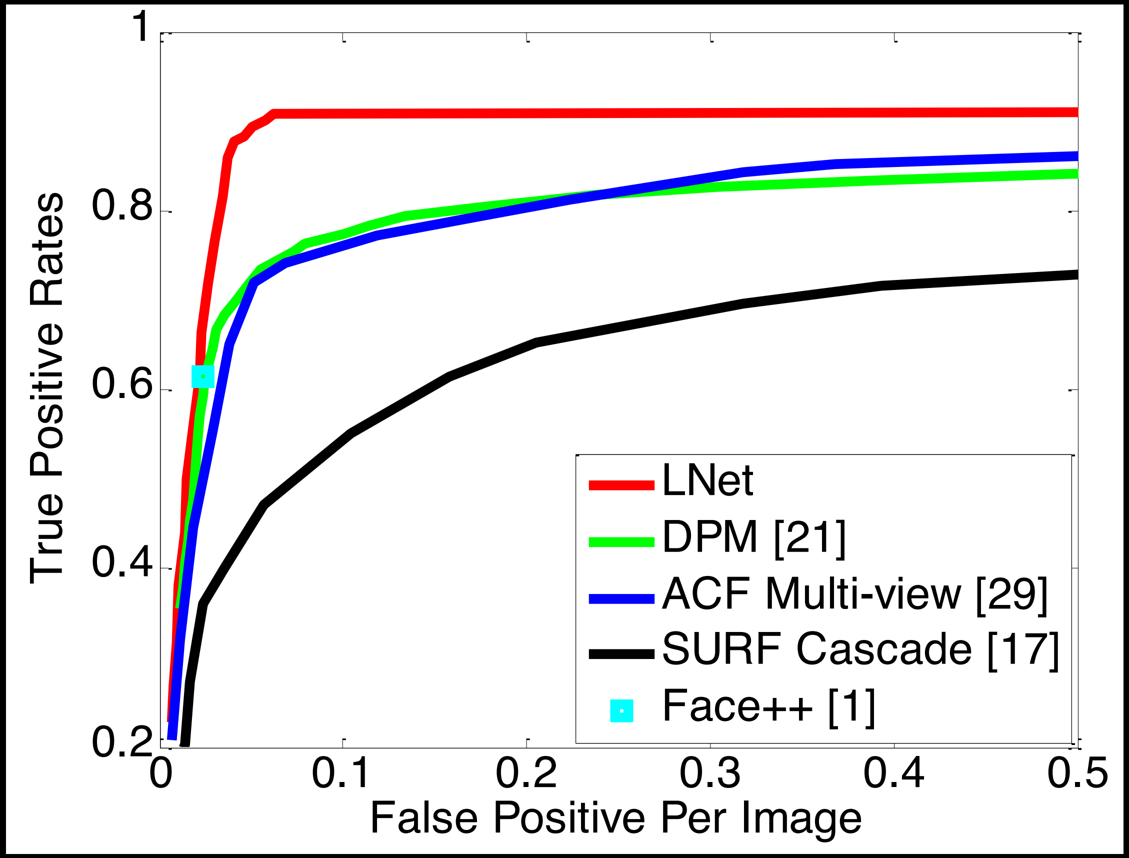




Deep Face Representation

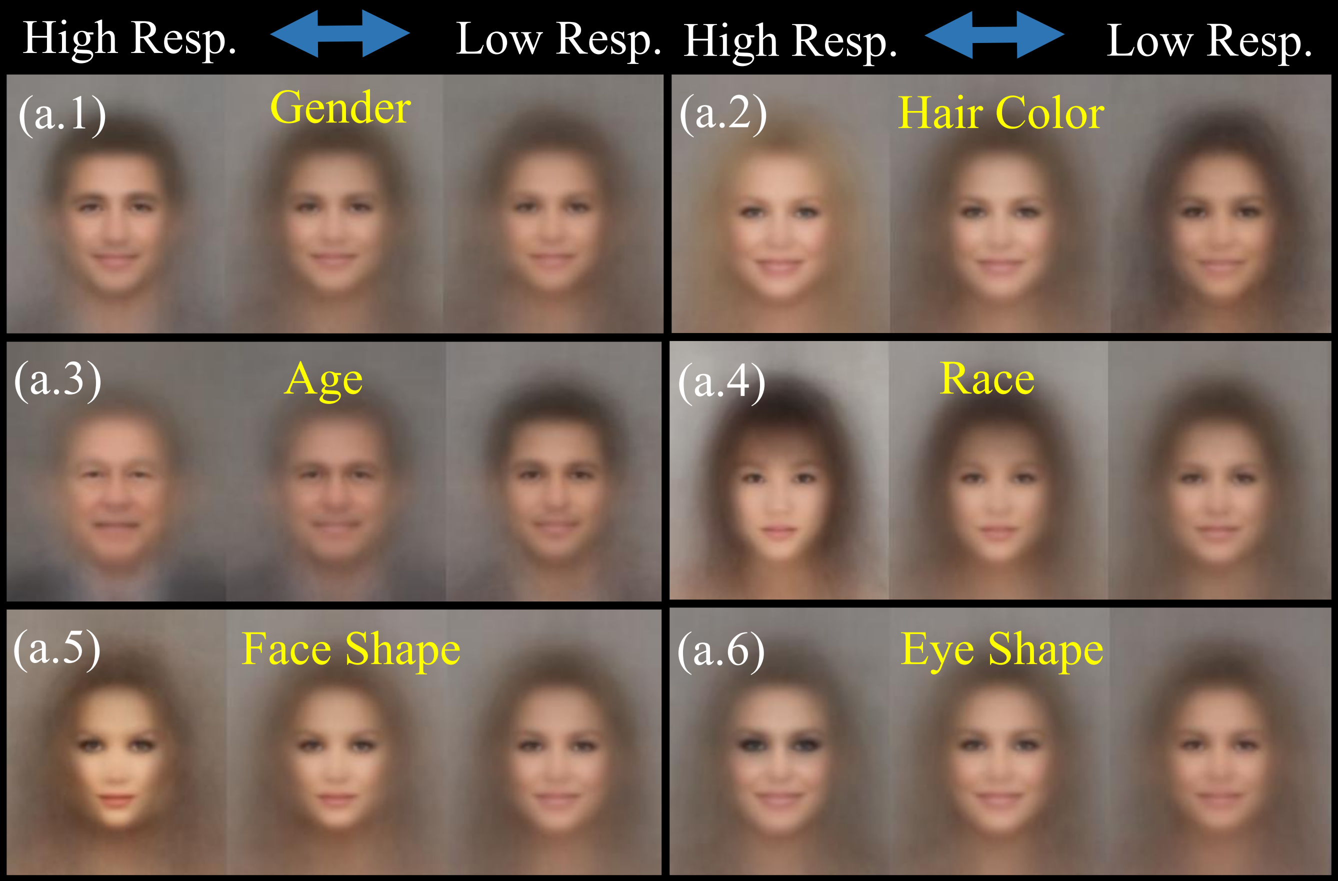
- Face localization performance on MobileFace





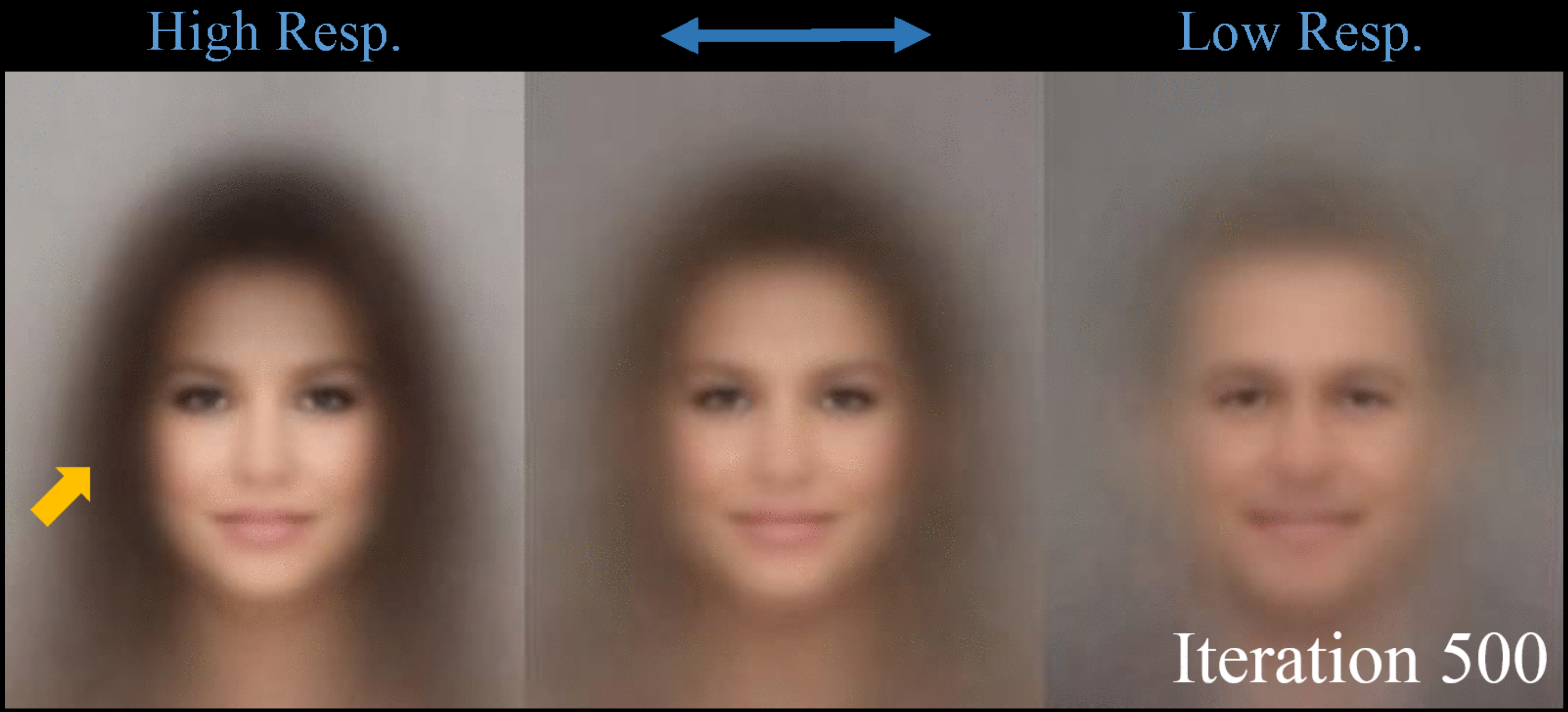
Deep Face Representation

- Pre-training with identities discovers semantic concepts



Deep Face Representation

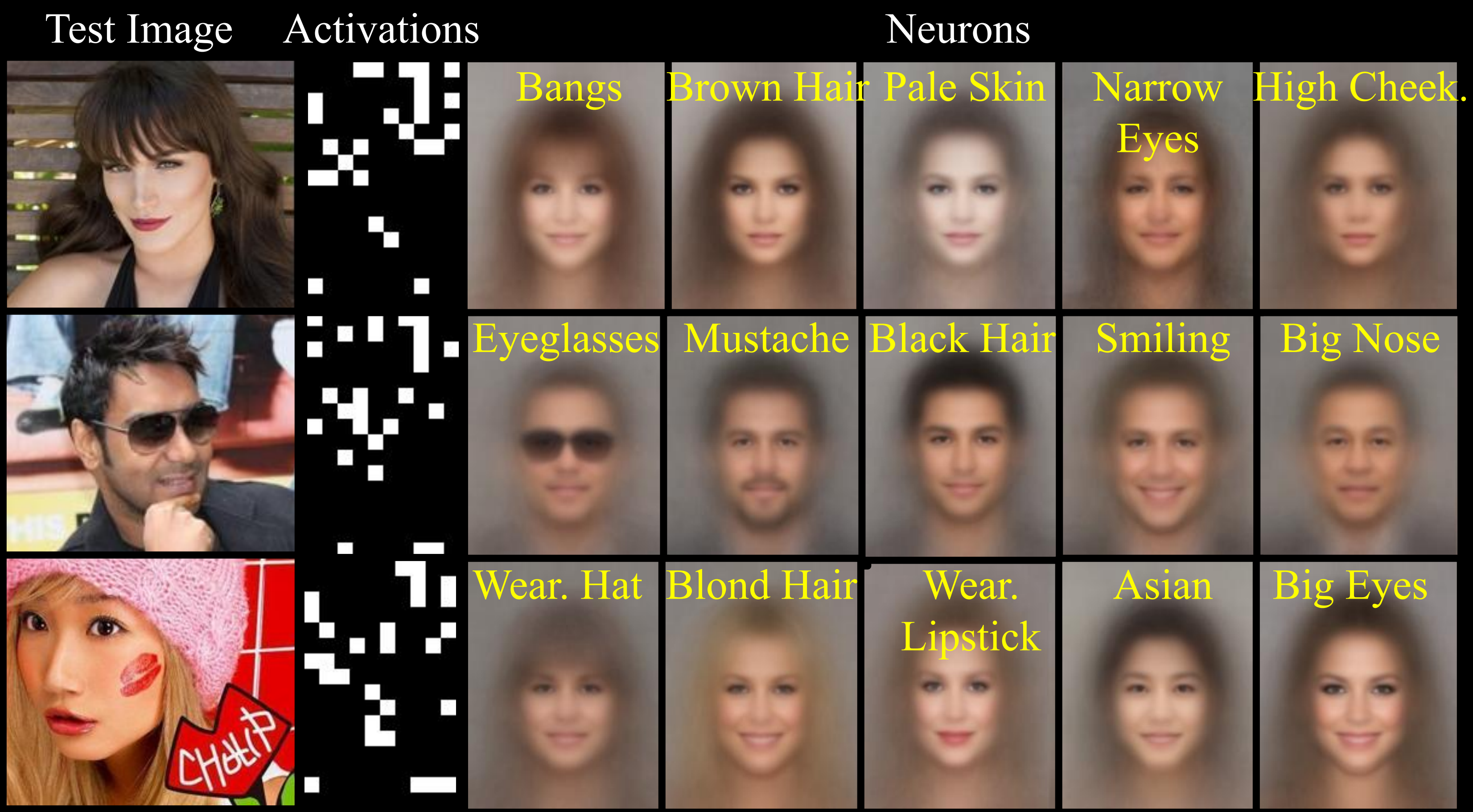
- Pre-training with identities discovers semantic concepts



Race (Neuron #131)

Deep Face Representation

- Fine-tuning with attributes expands semantic concepts



Deep Face Representation

- Fine-tuning with attributes expands semantic concepts

High Resp. Low Resp.

Thick Lip (Neuron #152)

Face Attributes Recognition

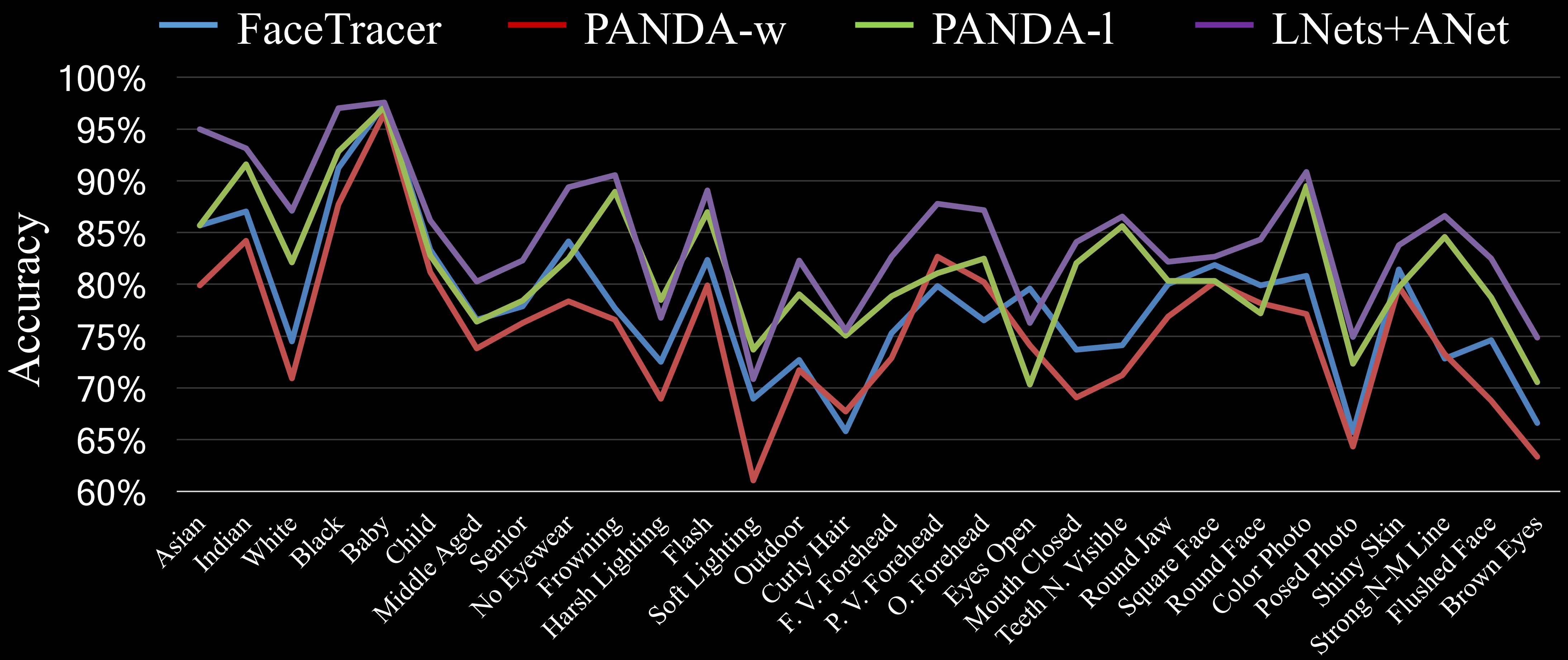
- Attribute recognition performance (40 attributes)

	CelebA (200K)	LFWA (13K)
FaceTracer	81%	74%
PANDA-w	79%	71%
PANDA-l	85%	81%
SC+ANet	83%	76%
LNets+ANet(w/o)	83%	79%
LNets+ANet	**87%**	**84%**

Running Time: LNets (35ms), ANet (14ms)

Face Attributes Recognition

- Performance on unseen attributes (30 attributes)



Part II: Deep Fashion Understanding

“DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations”, *CVPR 2016*

“Fashion Landmark Detection in the Wild”, *ECCV 2016*

Challenges

Face Variations

Cloth Variations

Overall Pipeline

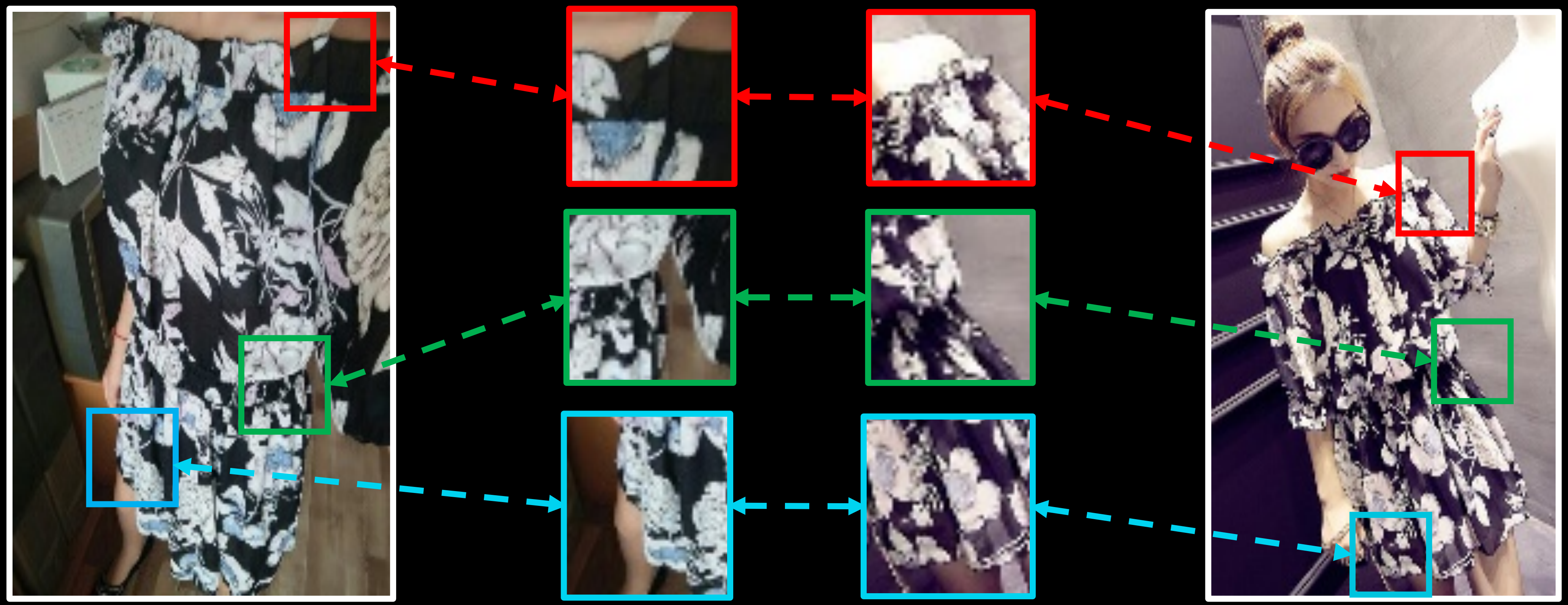
Clothes Detection

Overall Pipeline

Clothes Detection

Clothes Alignment

Overall Pipeline



Clothes Recognition

DeepFashion

- 800,000 images

- 50 categories
- 1,000 attributes
- 40,000 identities

- 1 bounding box
- 8 landmarks

Clothes Alignment

A set of fashion landmarks

Collars
Cuffs
Waistlines
Hemlines
…

(a.1)

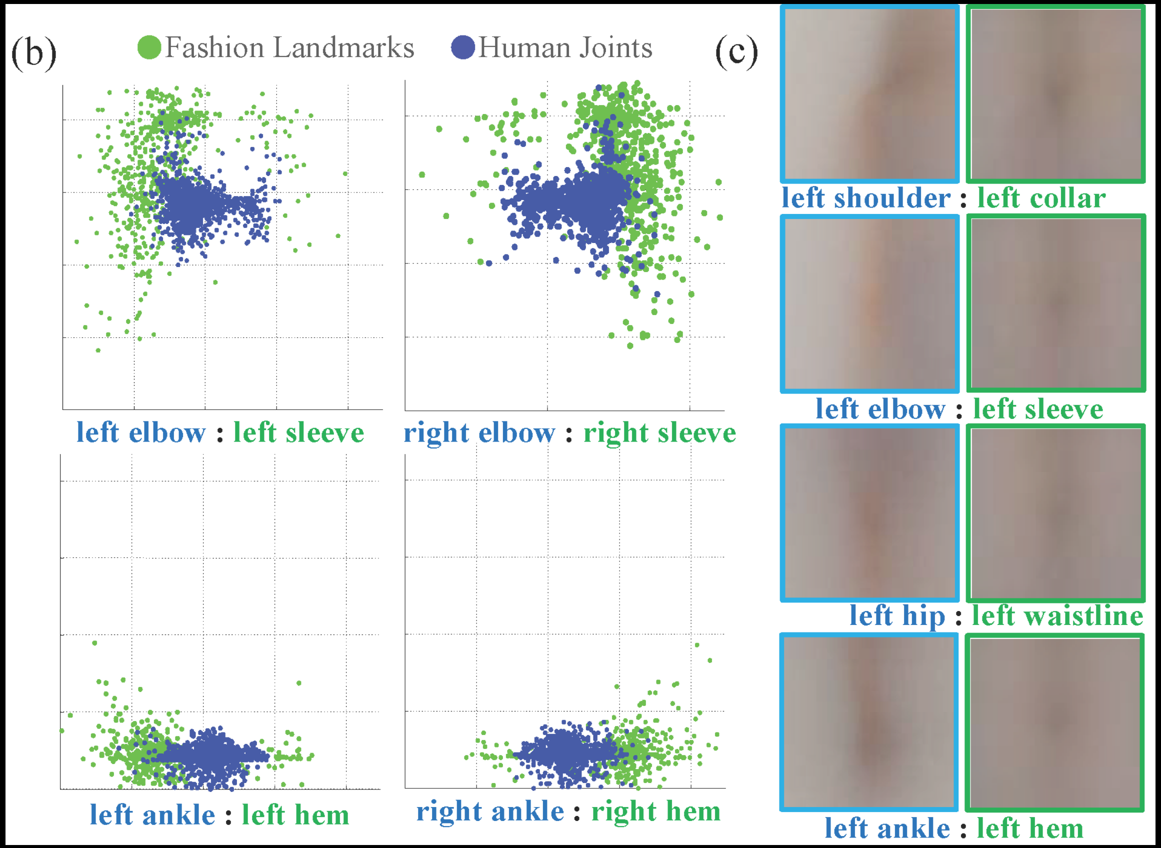
(a.2)

(a.3)

(a.4)

Clothes Alignment

More challenging than human pose estimation

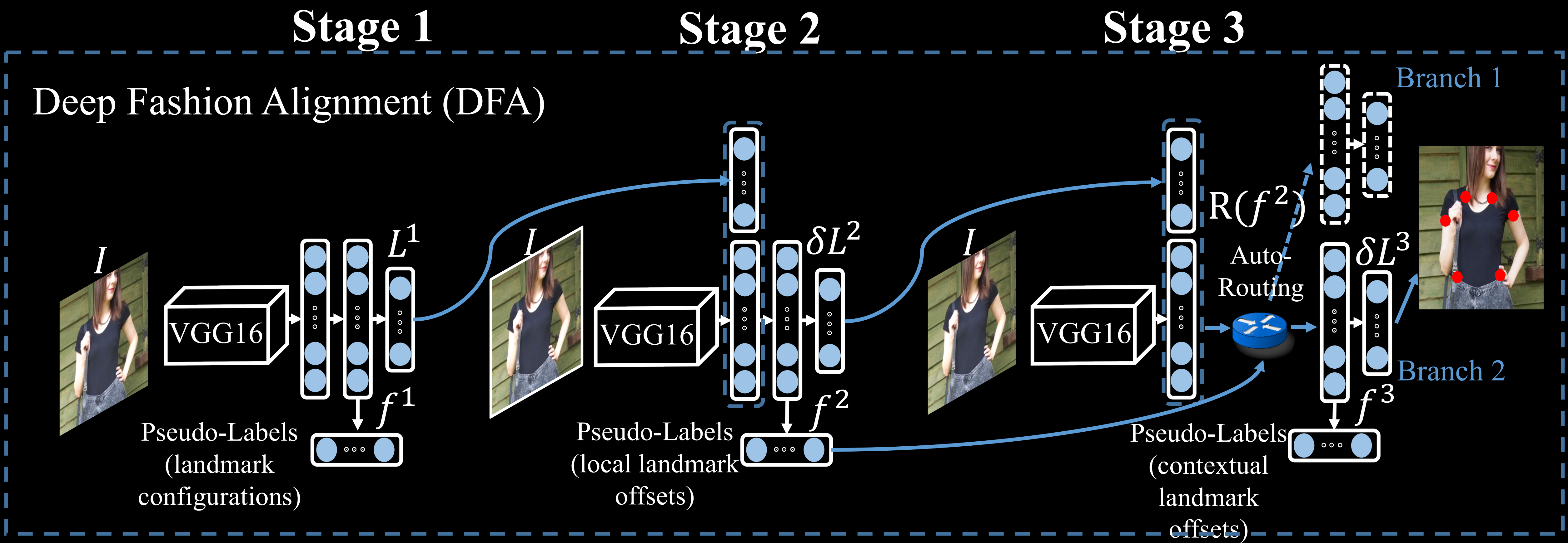


Geometry

Appearance

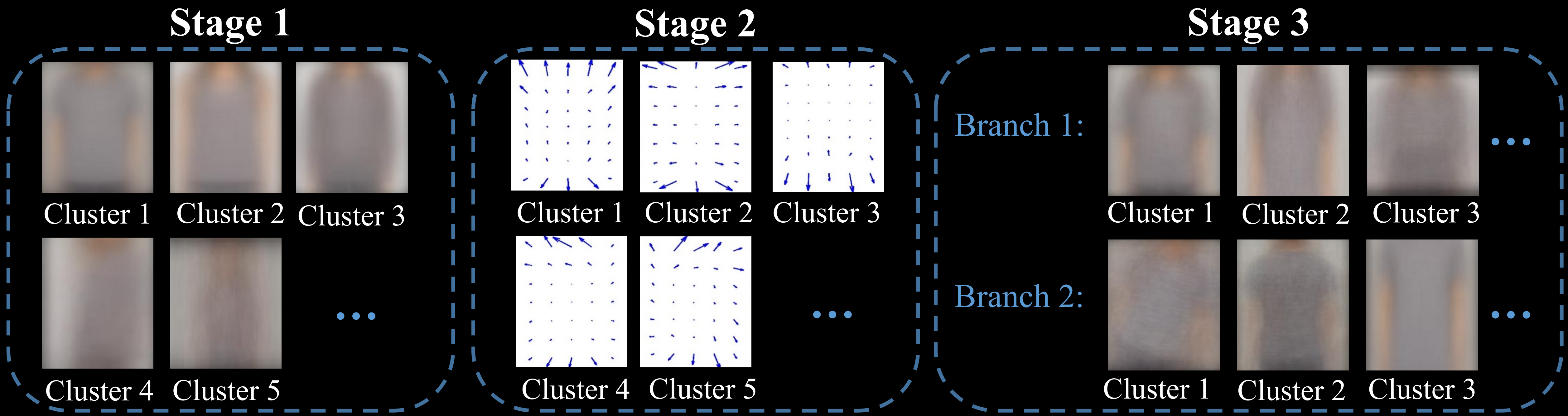
Clothes Alignment

Pipeline



Clothes Alignment

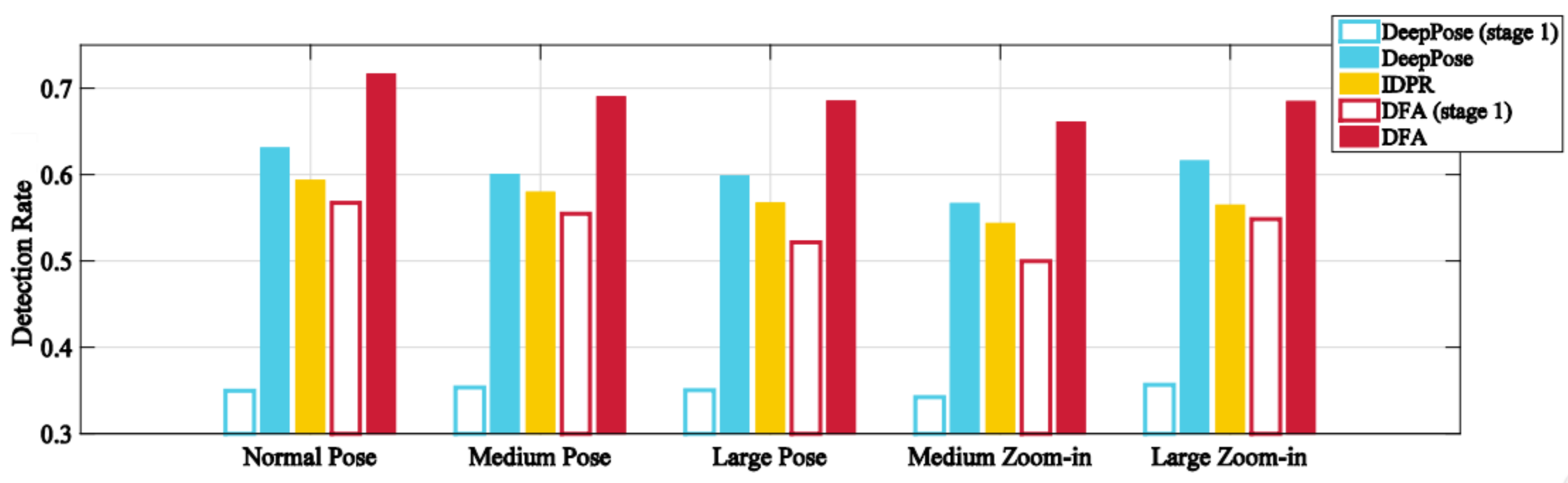
Reduce variations by pseudo-labels



Obtain codebook by k-means clustering in label space

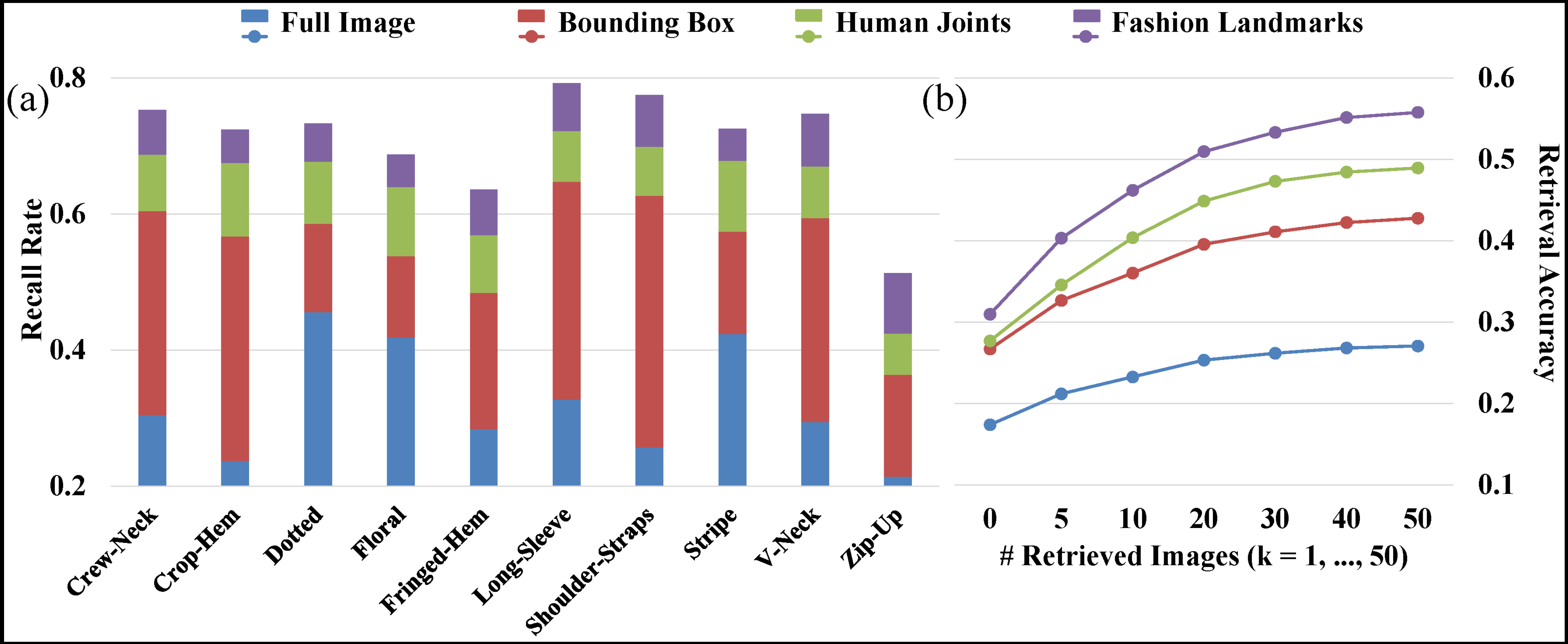
Clothes Alignment

Performance



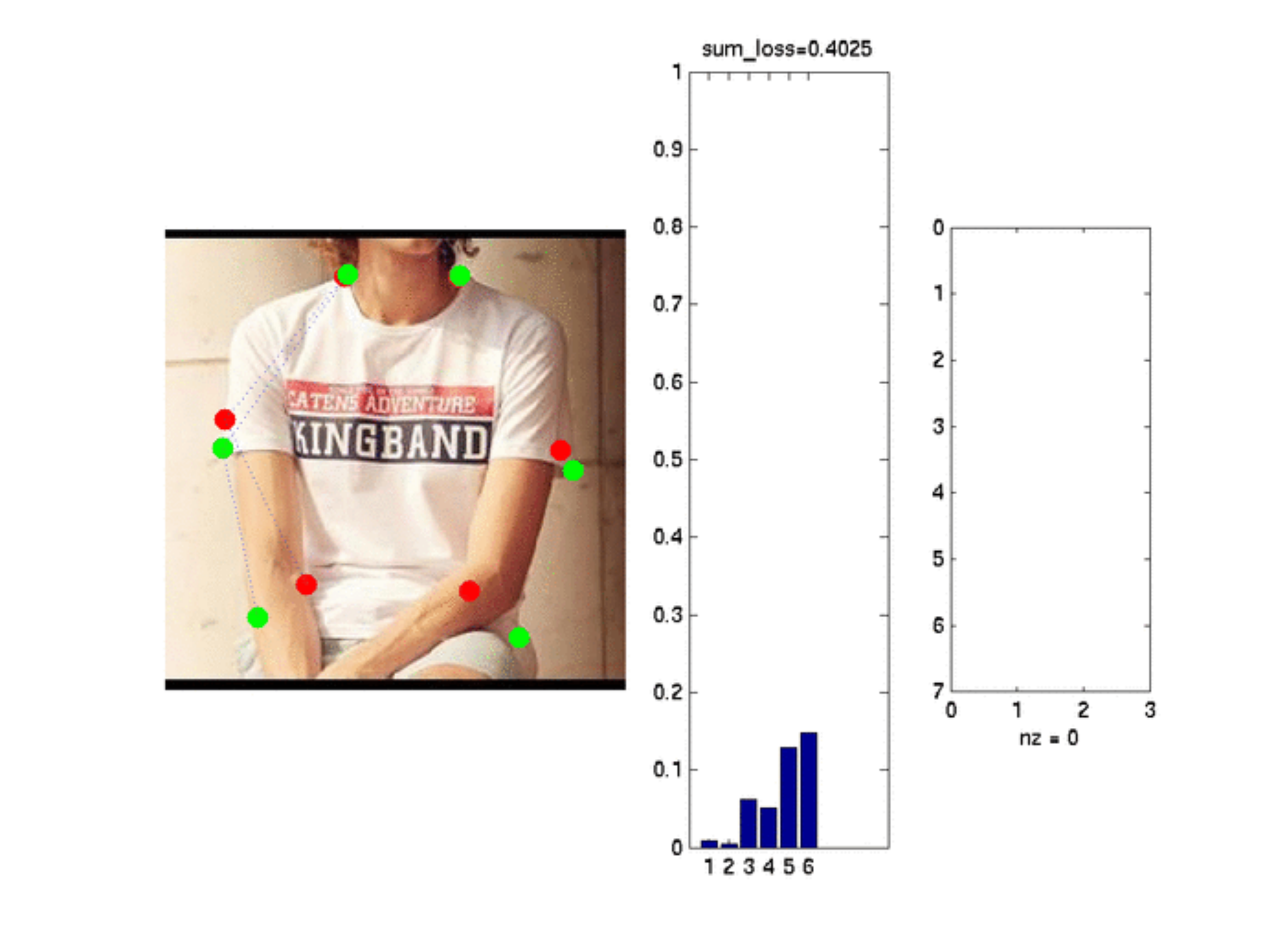
Clothes Alignment

More effective representation



Clothes Alignment

Demo



Clothes Recognition

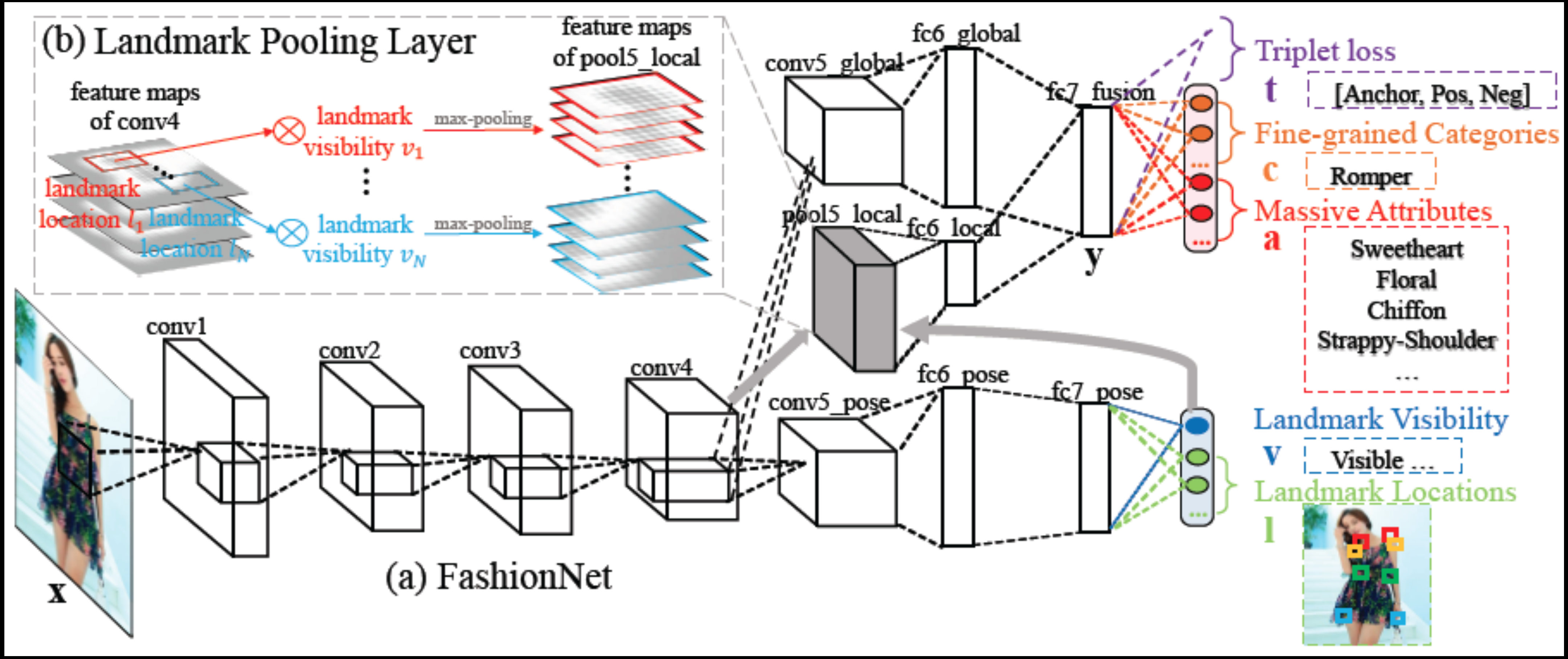
The interplay between identities and attributes

PID: 2000077658 (Forever21)

Ringer Tee (WOMEN)

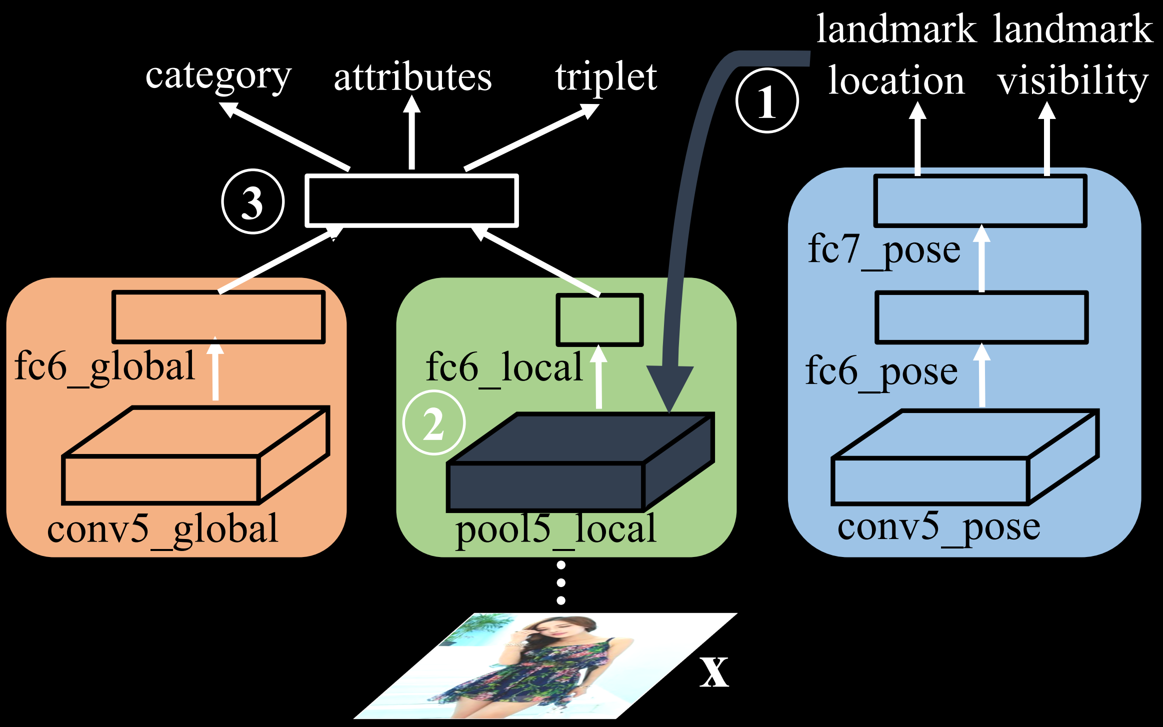
FashionNet

End-to-end System



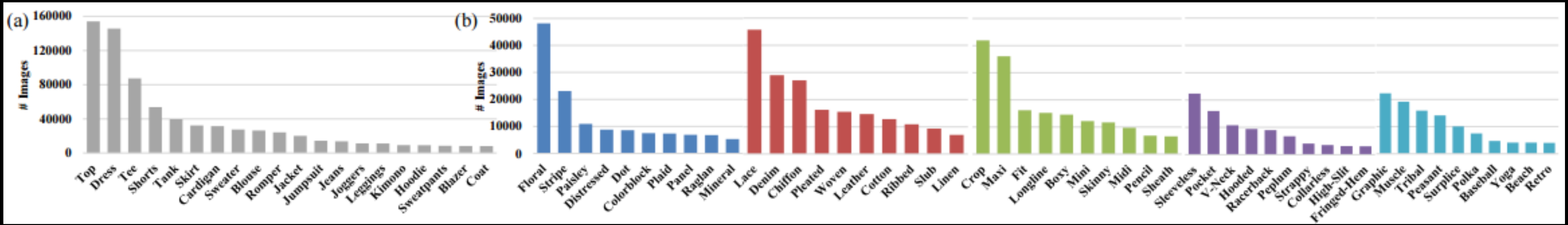
FashionNet

Forward/Backword Pass



Clothes Recognition

Attributes are noisy and imbalanced

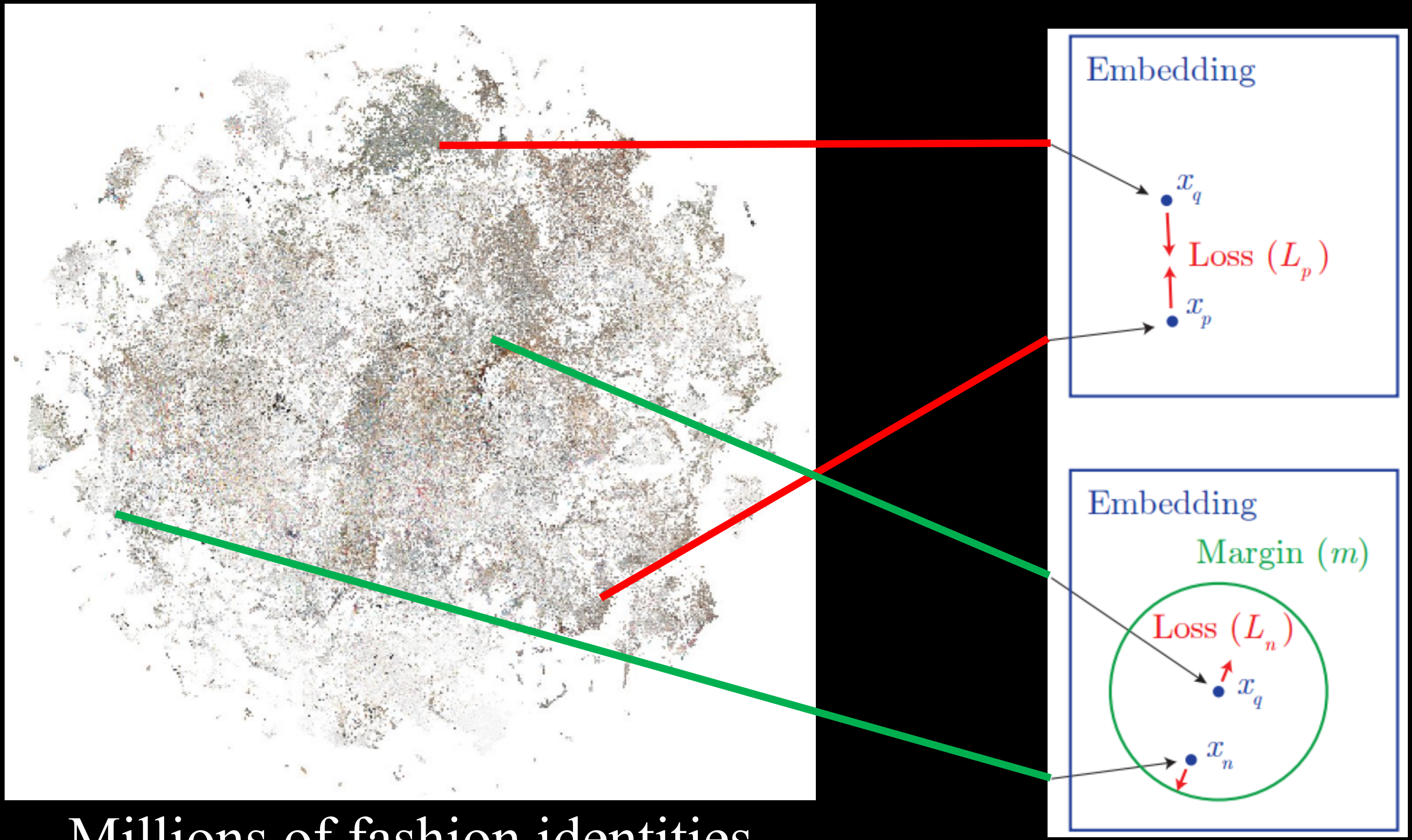


$$J = \sum_{i=1}^{n} \sum_{j=1}^{c_+} \sum_{k=1}^{c_-} \max(0, 1 - f_j(\boldsymbol{x}_i) + f_k(\boldsymbol{x}_i))$$

Multi-label Ranking Loss

Clothes Recognition

The number of identities are huge



Millions of fashion identities

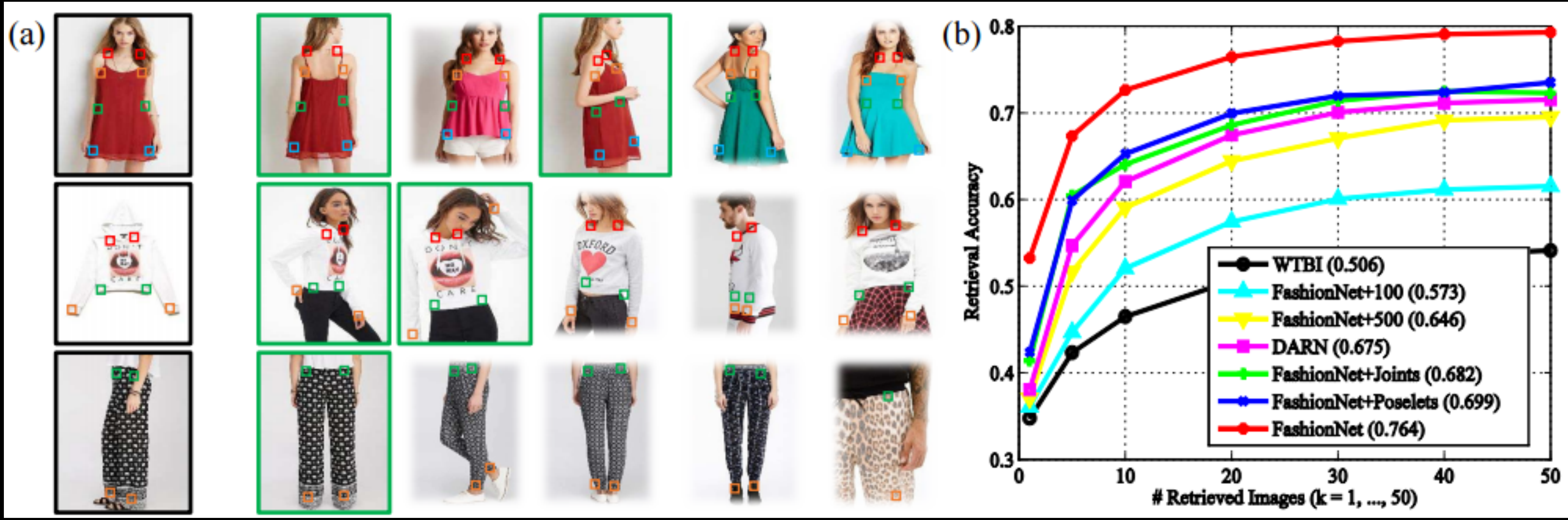
Positive Pair

Negative Pair

Hard Negative Mining

Clothes Recognition

In-shop Clothes Retrieval



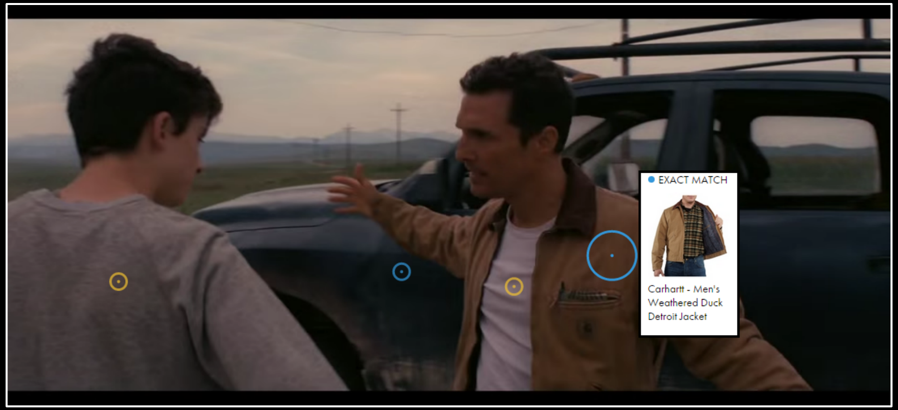
Clothes Recognition

Consumer-to-shop Clothes Retrieval

Clothes Recognition

Applications

Similar Style Retrieval



Cloth Spotting in Video

Street-to-shop

Fashion Assistant

Part III: Deep Scene Understanding

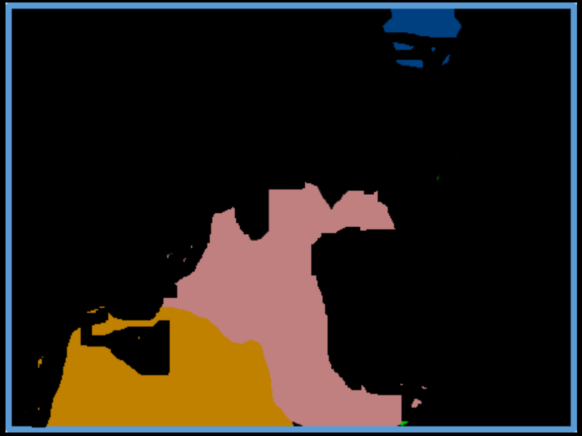
“Semantic Image Segmentation via Deep Parsing Network”, *ICCV 2015 (oral)*

“Not All Pixels Are Equal: Difficulty-aware Semantic Segmentation via Deep Layer Cascade”, *CVPR 2017 (spotlight)*

Problem

Problem

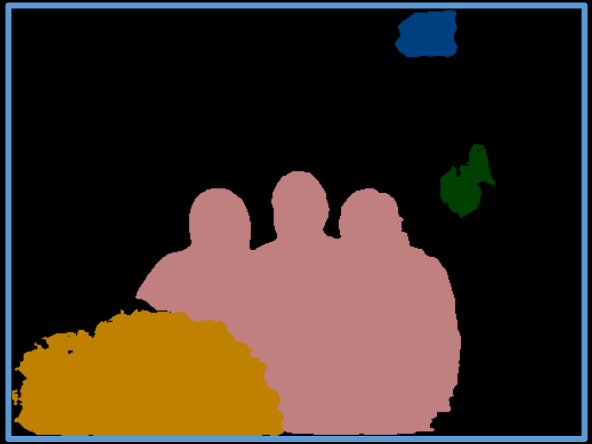
Previous Attempts



SVM

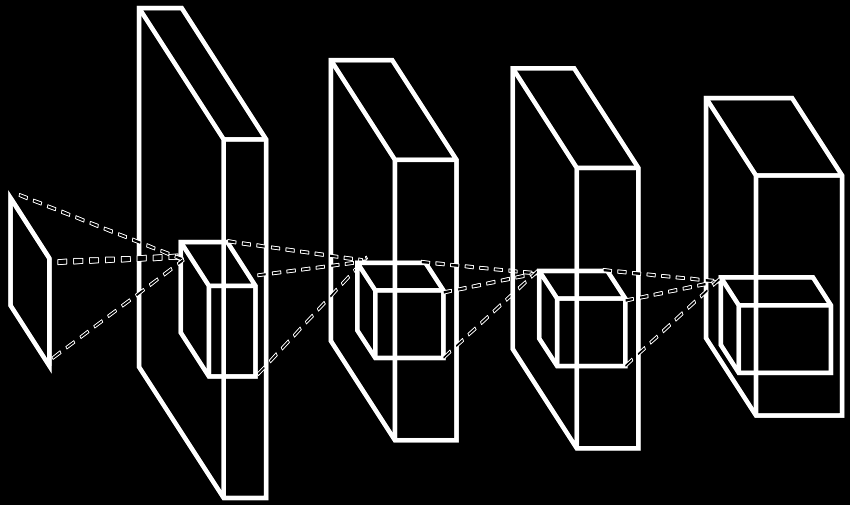
SVM + MRF

CNN



CNN + MRF ?

State-of-the-arts



Fully Convolutional Network
[Long et al. CVPR 2015]

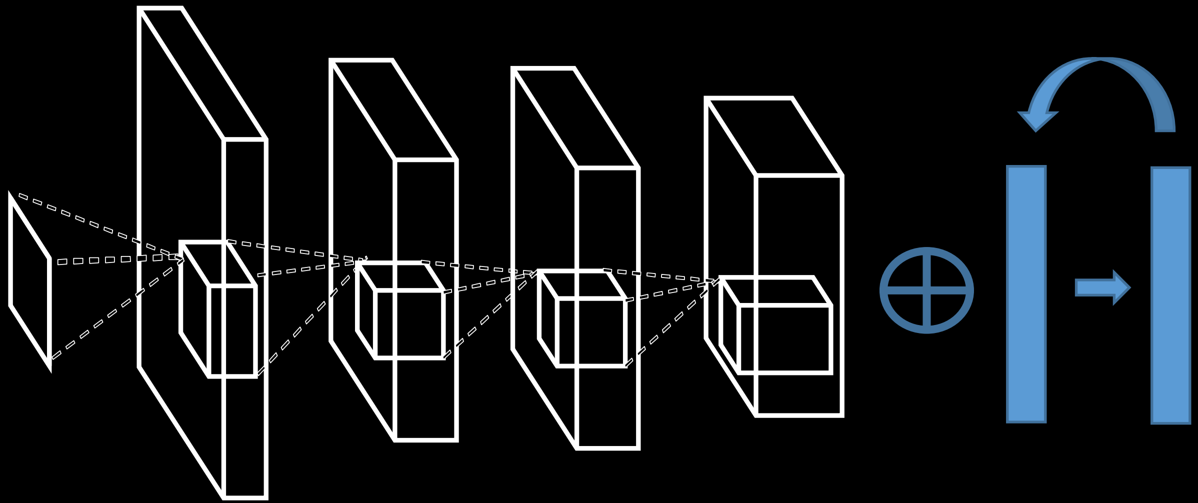
Learned Features	✓
Pairwise Relations	✗
Joint Training	-
# Iterations	-

State-of-the-arts

DeepLab
[Chen et al. ICLR 2015]

Learned Features	✓
Pairwise Relations	✓
Joint Training	✗
# Iterations	10

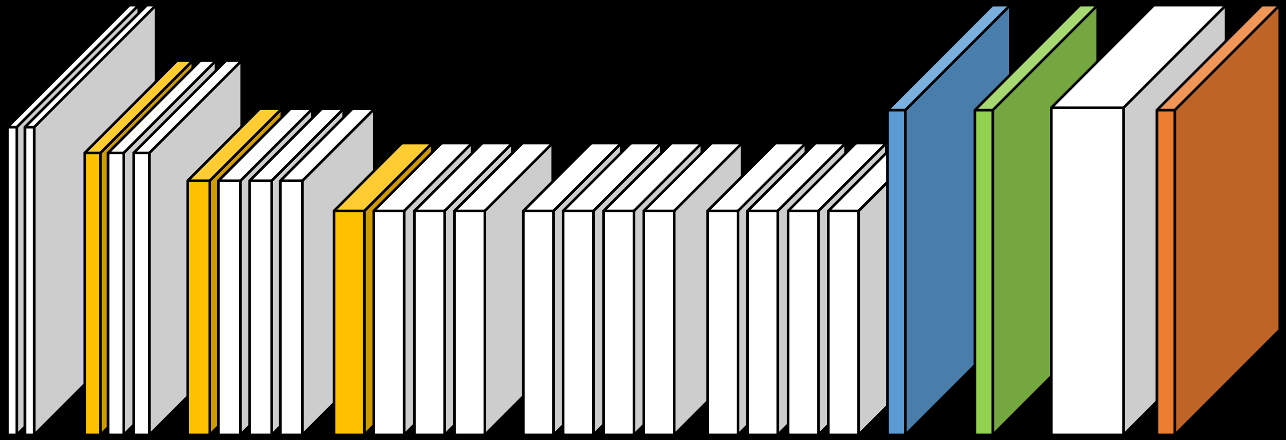
State-of-the-arts



CRF as RNN
[Zheng et al. ICCV 2015]

Learned Features	✓
Pairwise Relations	✓
Joint Training	✓
# Iterations	10

State-of-the-arts



Deep Parsing Network (DPN)

Learned Features	✓
Pairwise Relations	✓
Joint Training	✓
# Iterations	1

Contributions

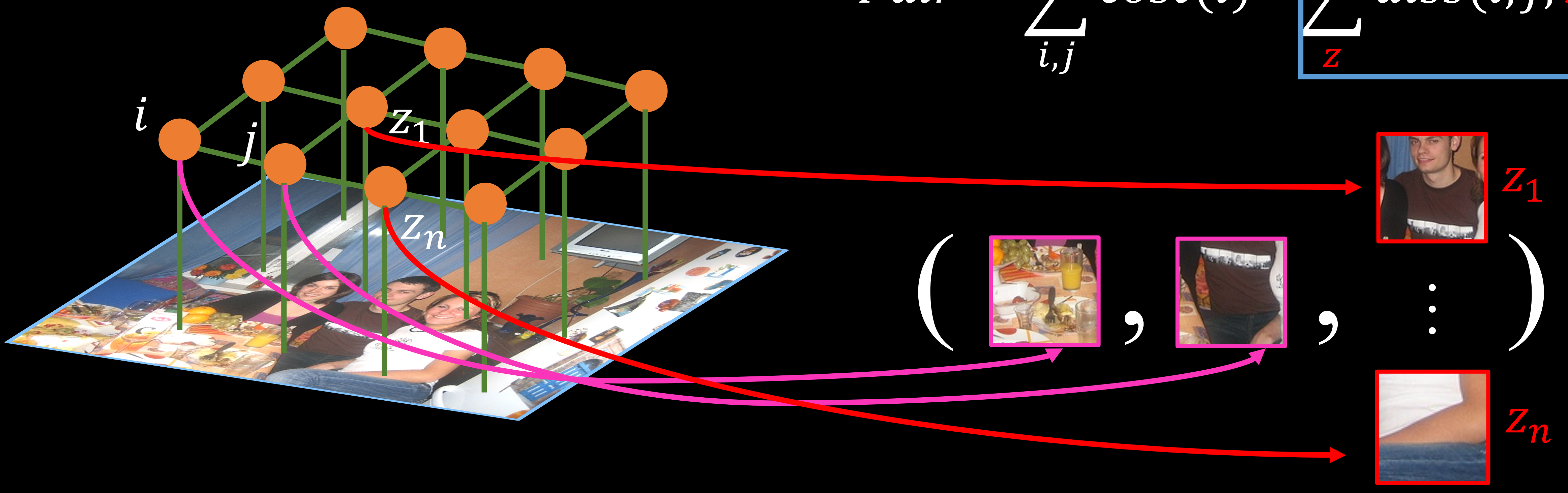
- Extend MRF to incorporate richer relationships
- Formulate mean field inference of high-order MRF as CNN
- Capable of joint training and one-pass inference

Richer Relationships in DPN

Triple Penalty

Pairwise Term

$$Pair = \sum_{i,j} cost(i) * \boxed{\sum_{z} diss(i,j;z)}$$



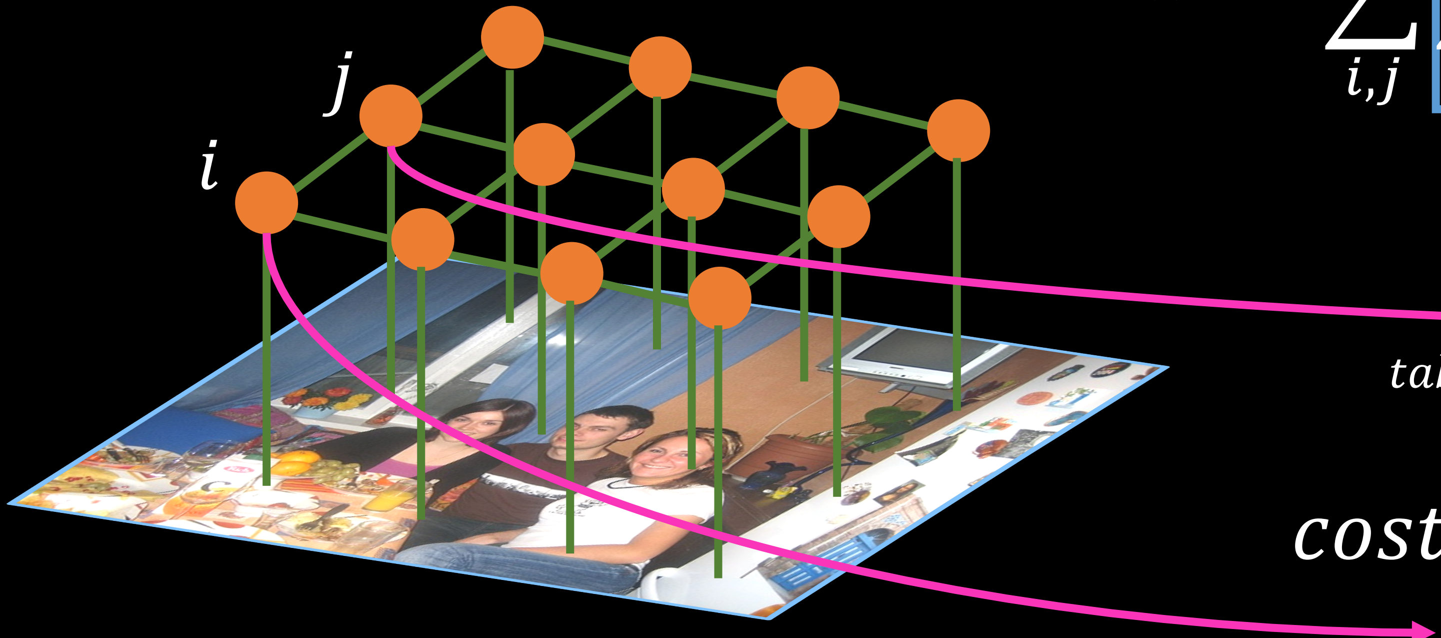
Triple Penalty

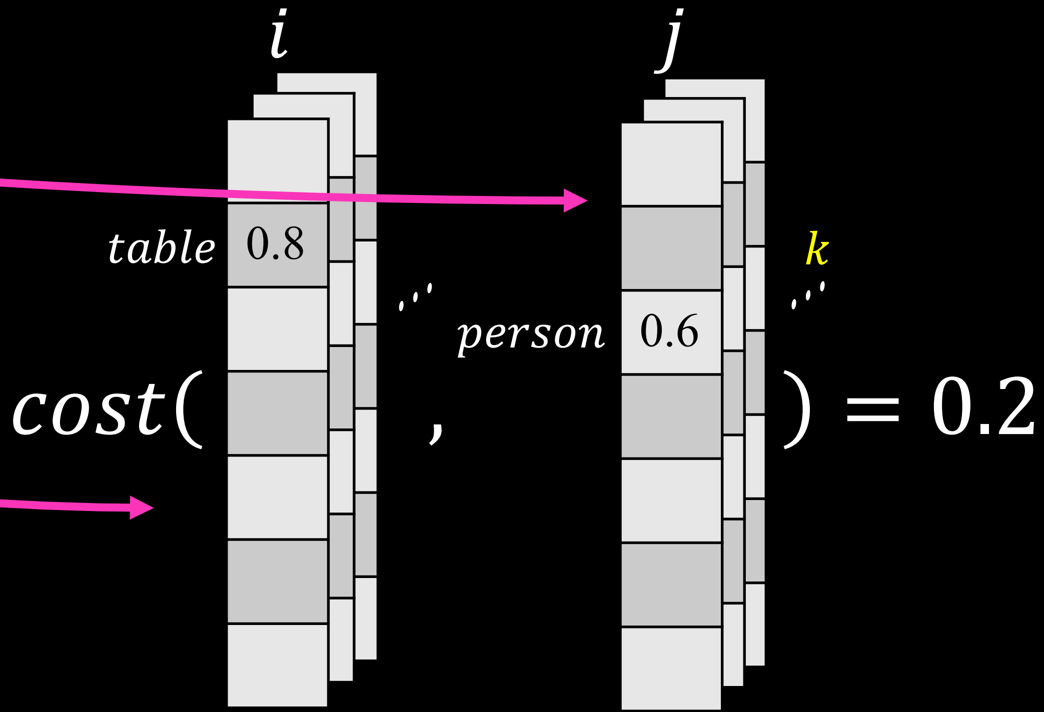
Richer Relationships in DPN

Mixture of Label Contexts

Pairwise Term

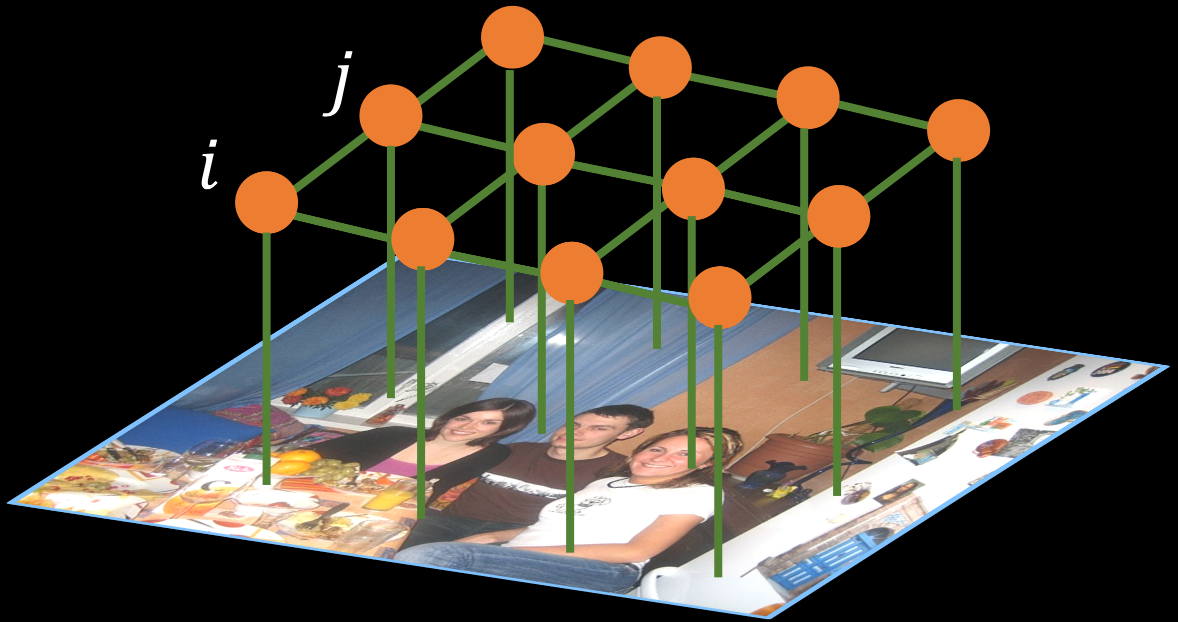
$$Pair = \sum_{i,j} \boxed{\sum_{k} cost_k(i,j)} * \sum_{z} diss(i,j;z)$$





Mixture of Label Contexts

Solve High-order MRF as Convolution



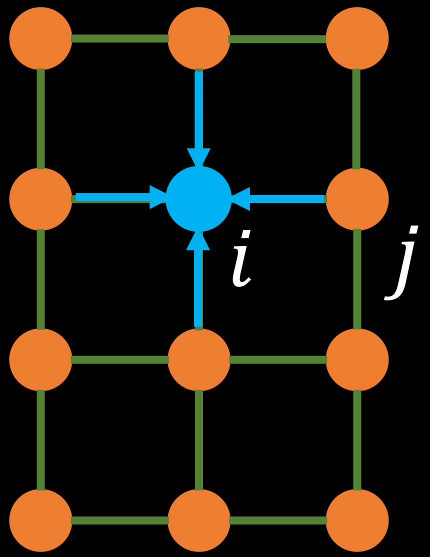
Pairwise Term

$$Pair = \sum_{i,j} \sum_{k} cost_k(i,j) * \sum_{z} diss(i,j;z)$$

Mean Field Solver

$$p_i \propto exp\left\{-\left(Unary_i + \sum_{j} Pair_{i,j} * p_j\right)\right\}$$

Solve High-order MRF as Convolution

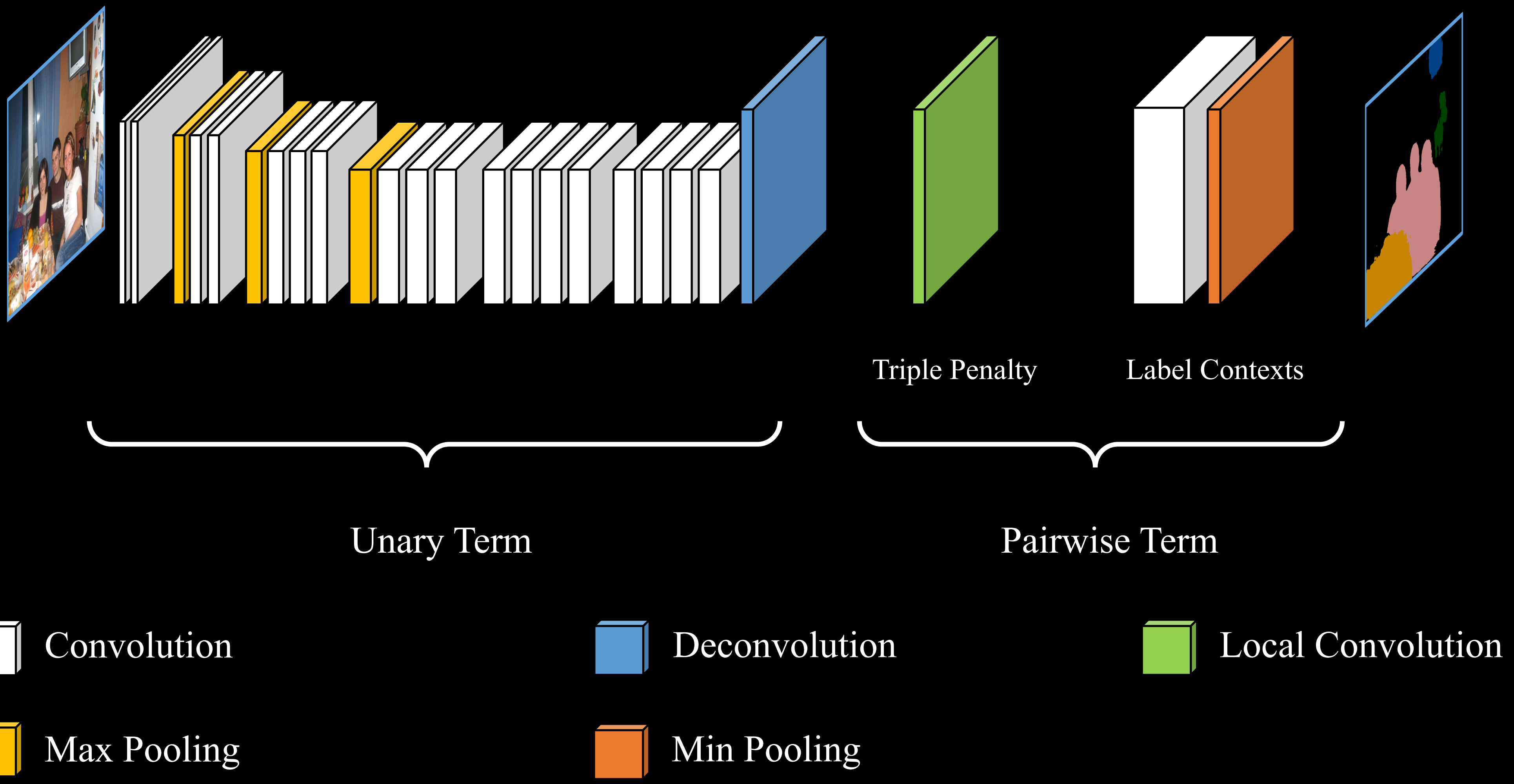


Iterative Updating Formula

$$p_i \propto exp\left\{-\left(\underbrace{\boxed{Unary_i}}_{\text{Summation}} + \underbrace{\boxed{\sum_j Pair_{i,j} * p_j}}_{\text{Convolution}}\right)\right\}$$

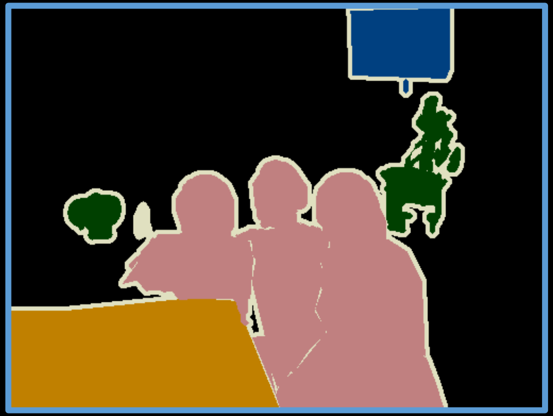
$Pair_{i,j}$: Different Types of Local and Global Filters

Deep Parsing Network

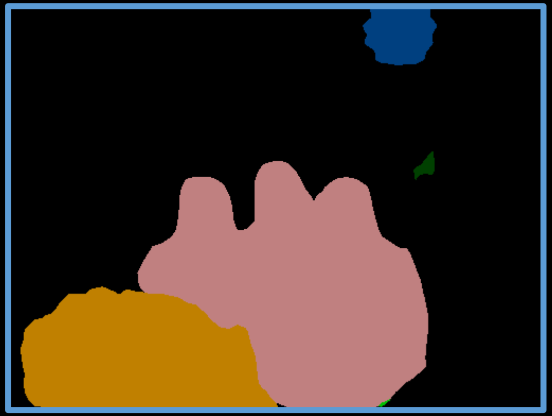


Deep Parsing Network

Original Image

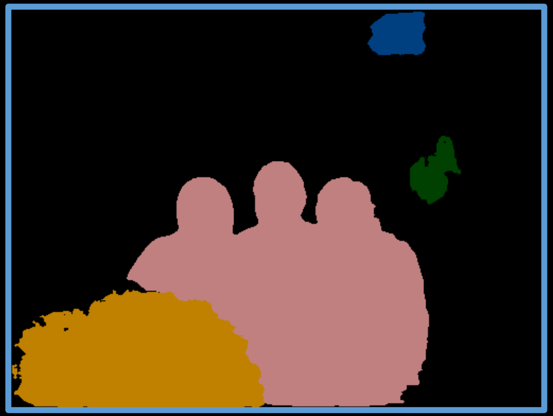


Ground Truth



Unary Term

Triple Penalty



Label Contexts

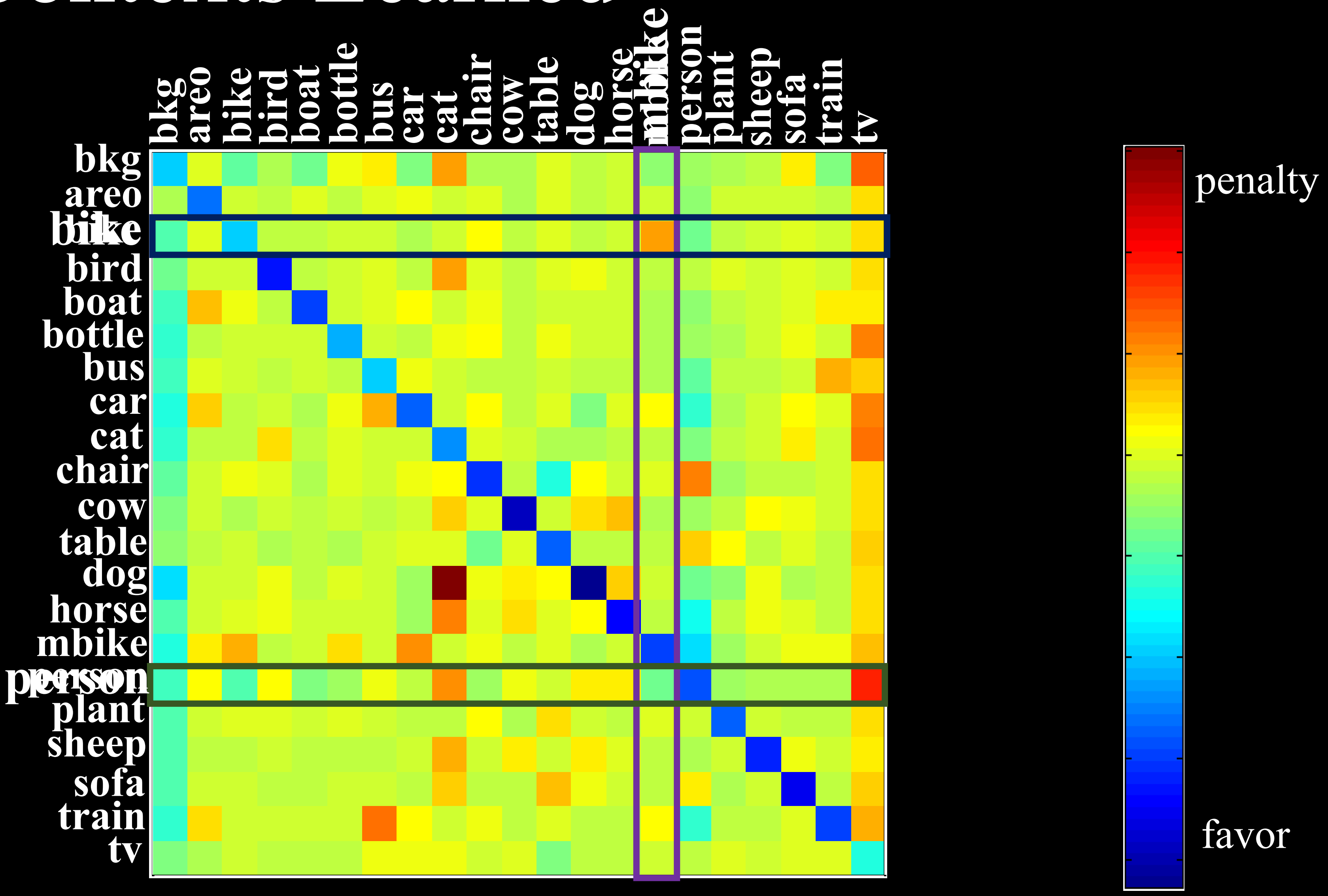
Joint Tuning

Overall Performance (Published Results)

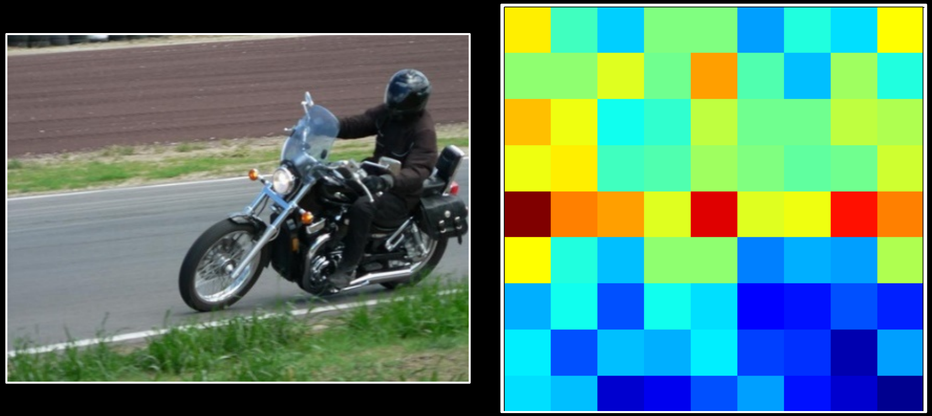
FCN	62.2
DeepLab†	73.9
CRFasRNN†	74.7
BoxSup†	75.2
DPN†	**77.5**

(PASCAL VOC 2012 Challenge test set)

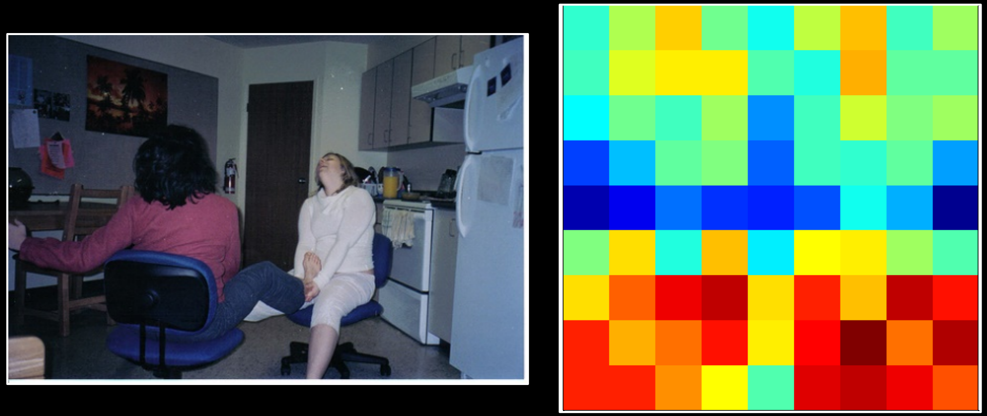
Label Contexts Learned



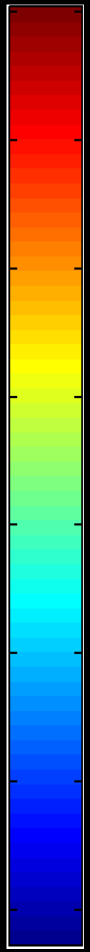
Label Contexts Learned



person : mbike



chair : person



Challenging Case

Problem

Input Video

State-of-the-art Method (4 FPS)

Deep Layer Cascade (17 FPS)

State-of-the-art

State-of-the-art Method (4 FPS)

- Why Slow?
  - Very Deep Backbone Network
  - High Resolution Feature Map

Fully Convolutional Network

Motivation

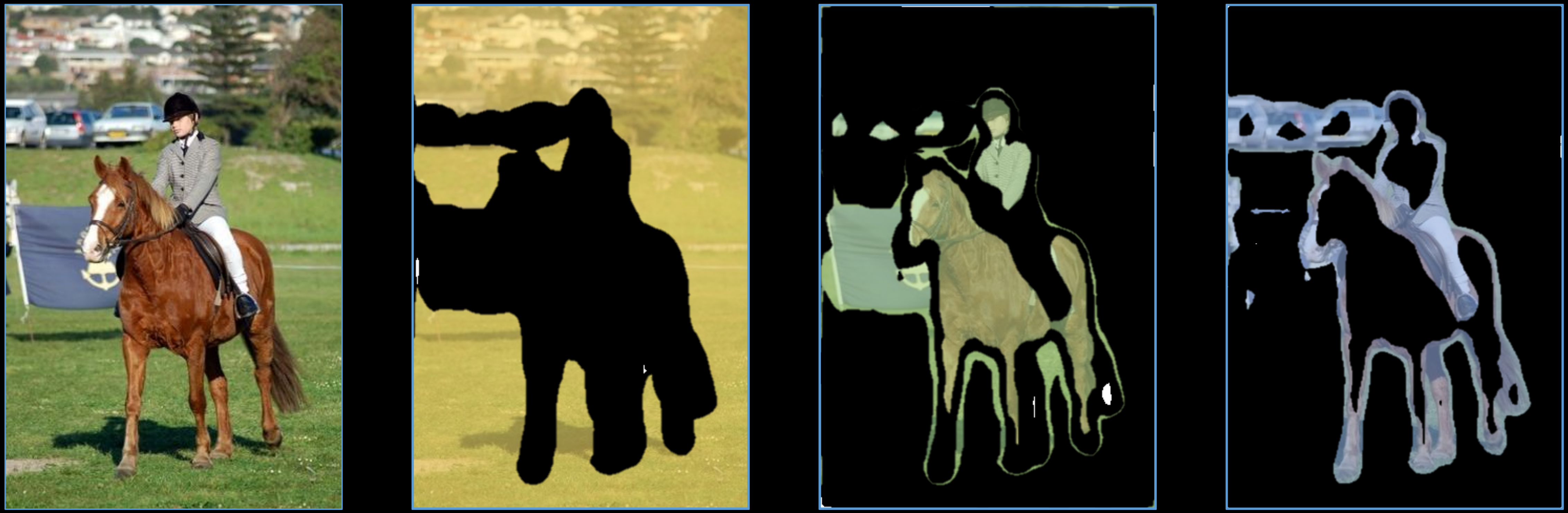
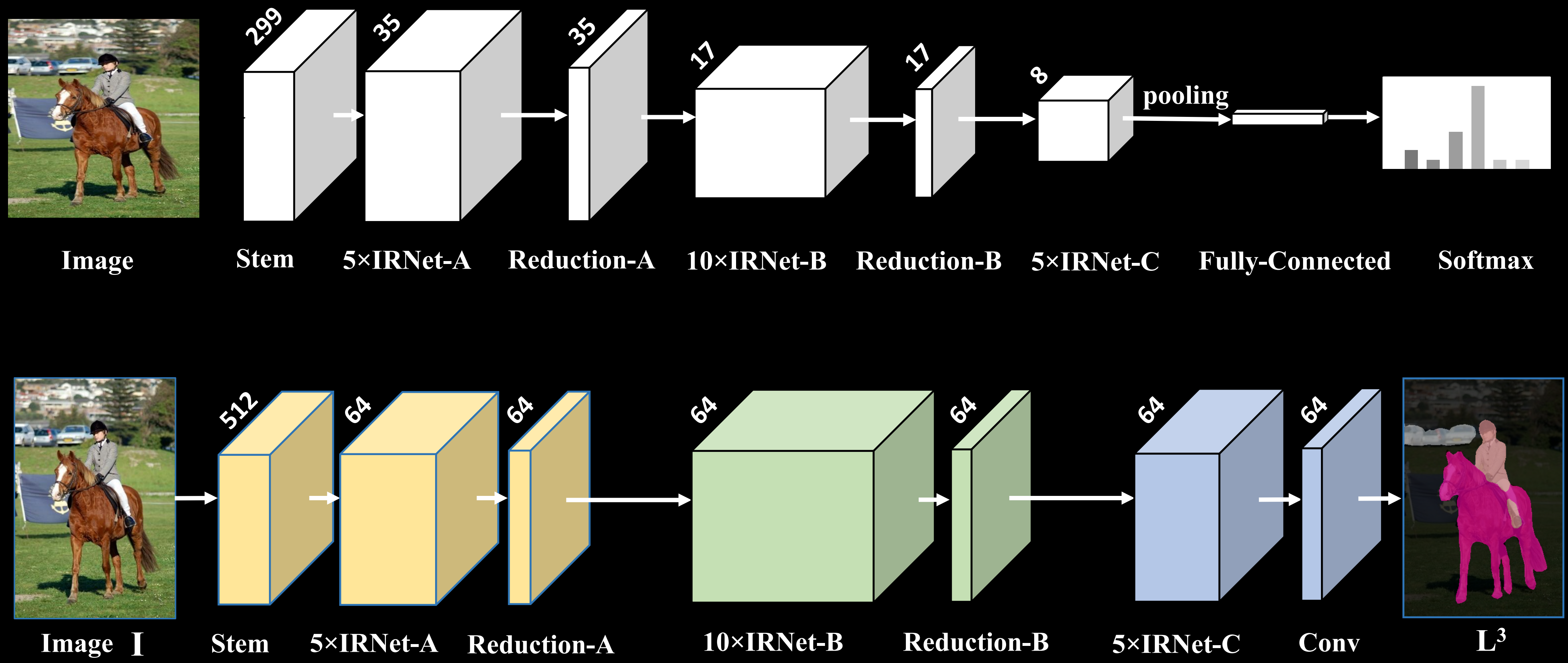
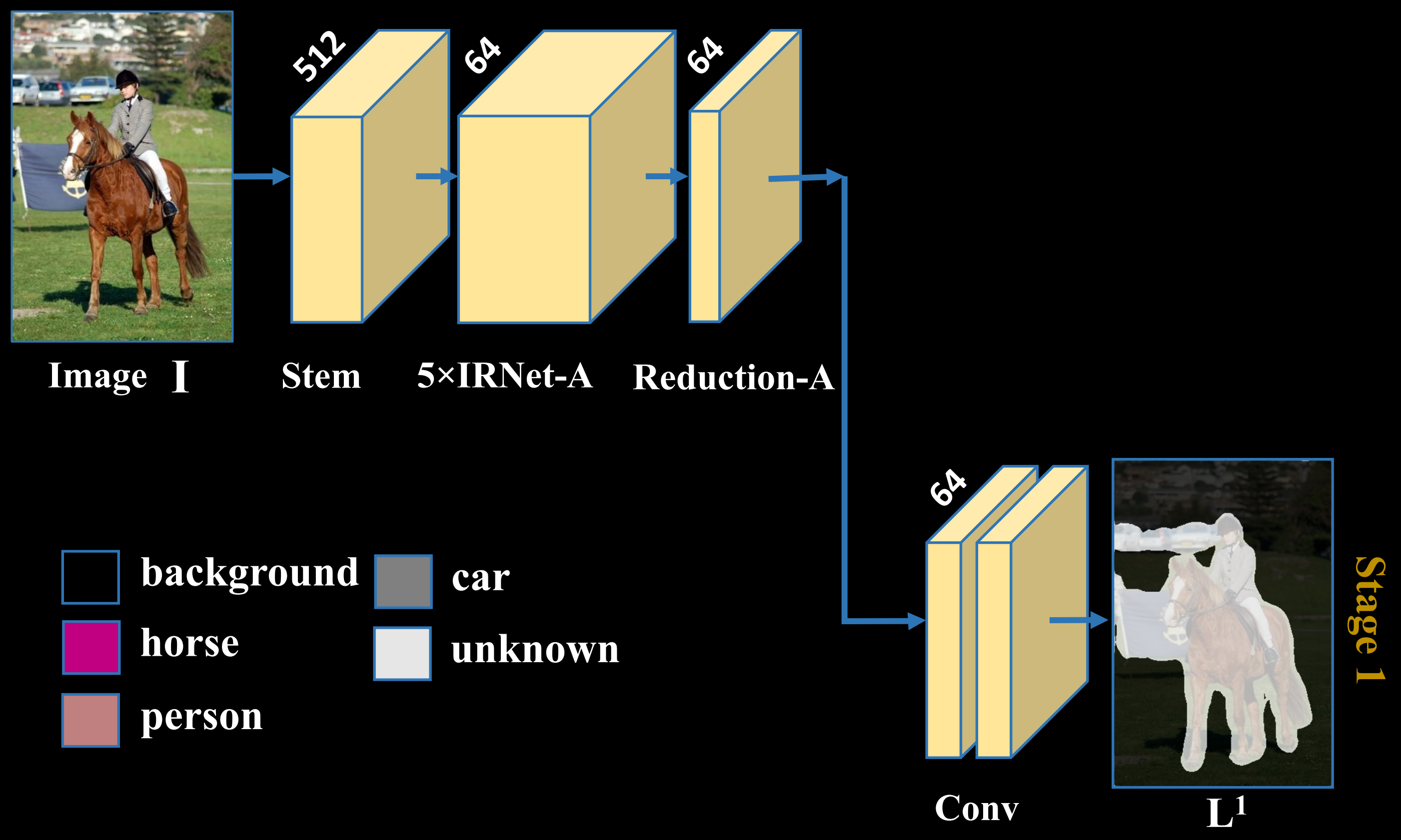


Image **Easy Region** **Moderate Region** **Hard Region**

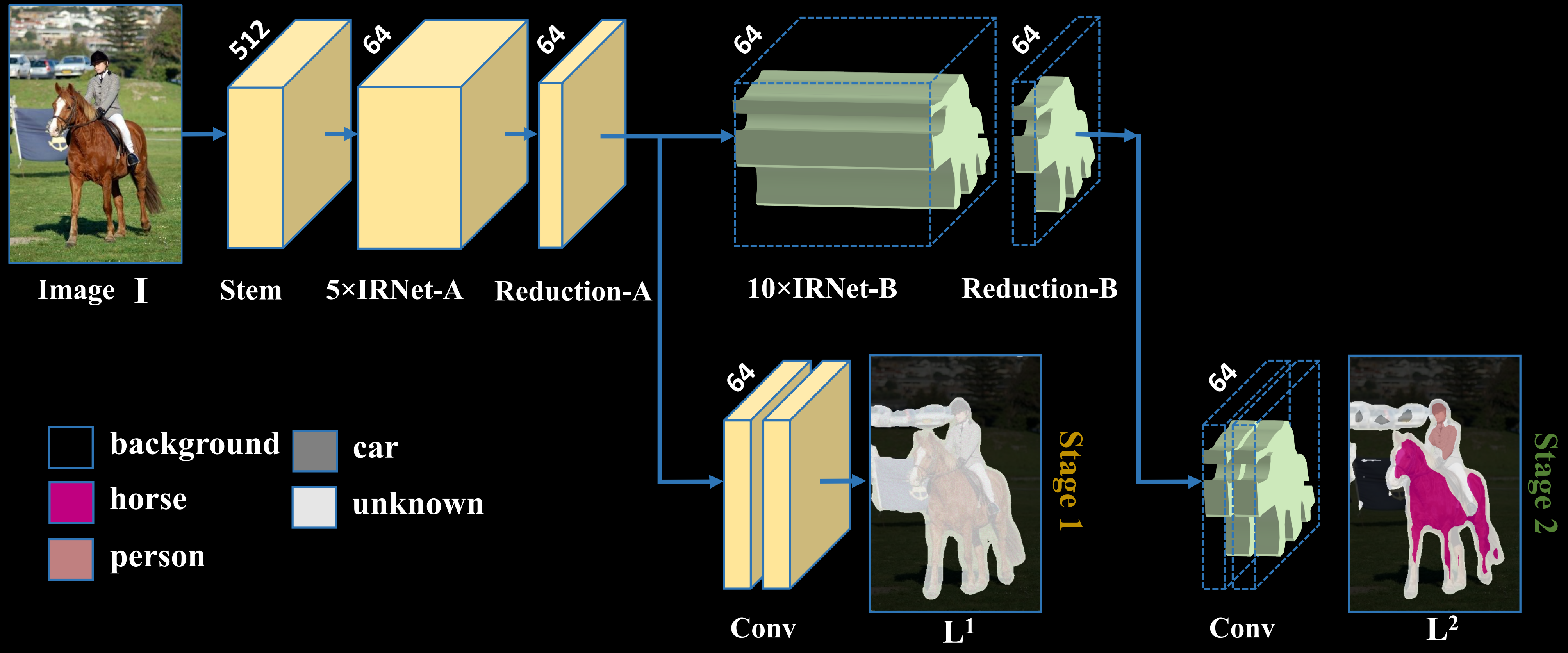
Contemporary Model



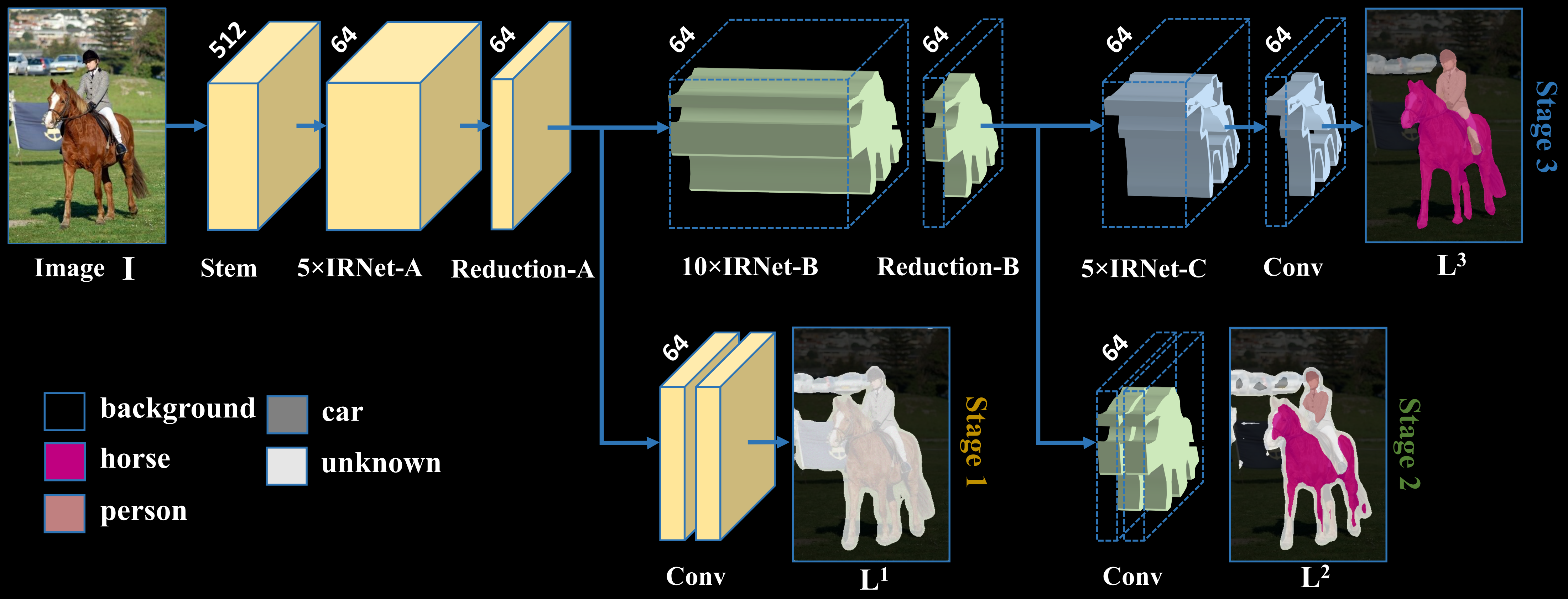
Deep Layer Cascade



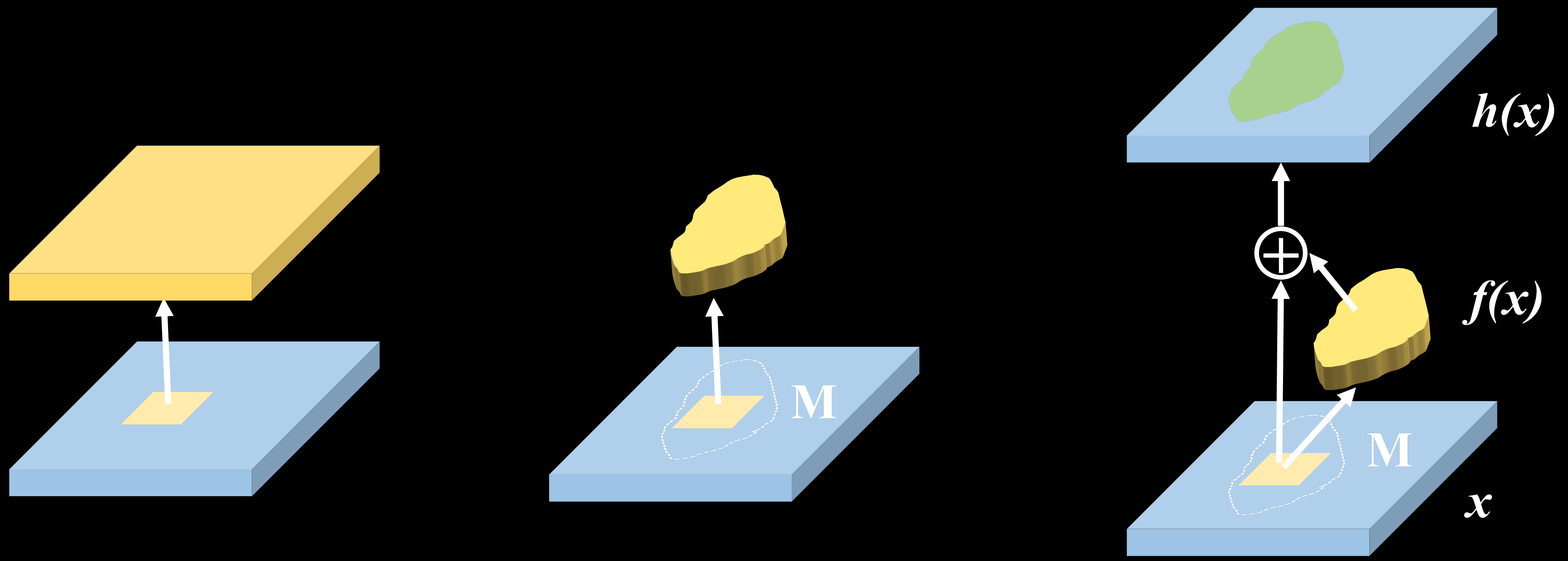
Deep Layer Cascade



Deep Layer Cascade



Region Convolution



Convolution **Region Convolution** **Region Convolution with Residual**

Performance

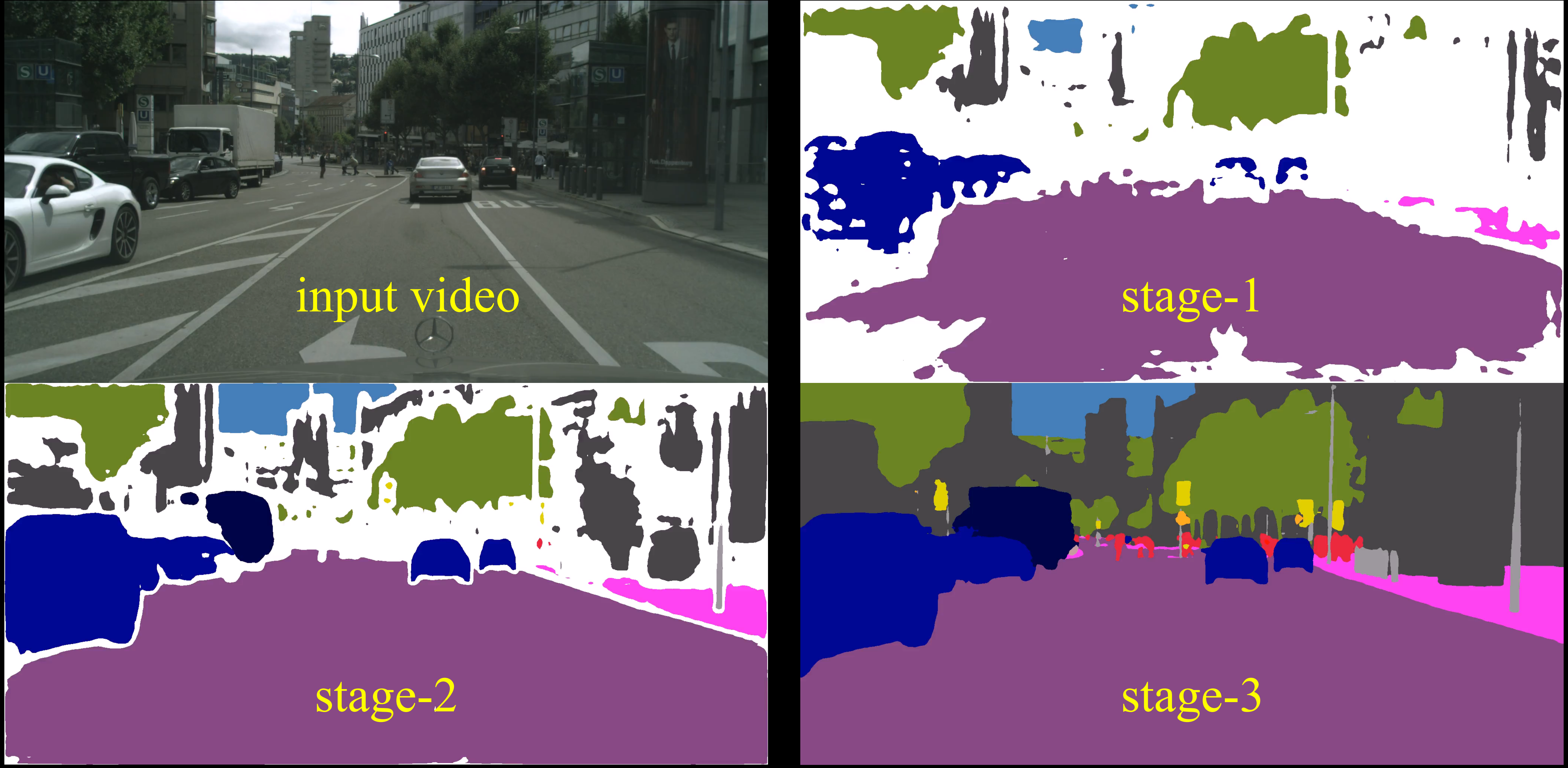
PASCAL VOC 2012

	mIoU	**FPS (Backbone Network)**
DPN	77.5	5.7
Adelaide	79.1	-
Deeplab-v2	79.7	7.1
LC(w/o COCO)	**80.3**	**14.7**
LC(with COCO)	**82.7**	

(PASCAL VOC 2012 Challenge test set)

Stage Visualization

(a) input image (b) stage-1 (c) stage-2 (d) stage-3 (e) ground truth



- Difficulty-Aware Learning Paradigm
- Region Convolution → Real-Time
- End-To-End Trainable Framework

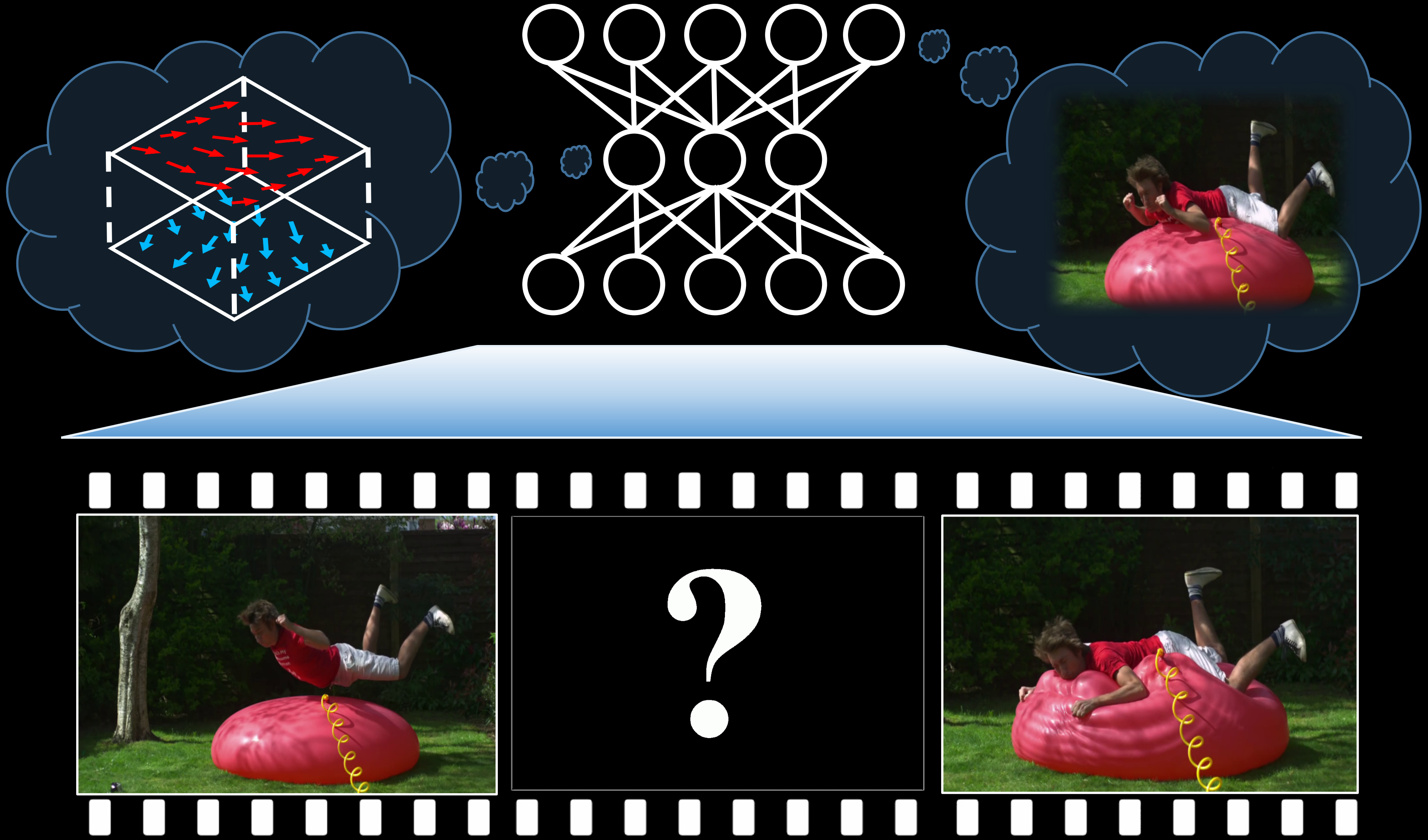
Part IV: Deep Motion Understanding

"Video Frame Synthesis using Deep Voxel Flow", *ICCV 2017 (oral)*

Video Frame Synthesis

- Problem

Video interpolation/ extrapolation

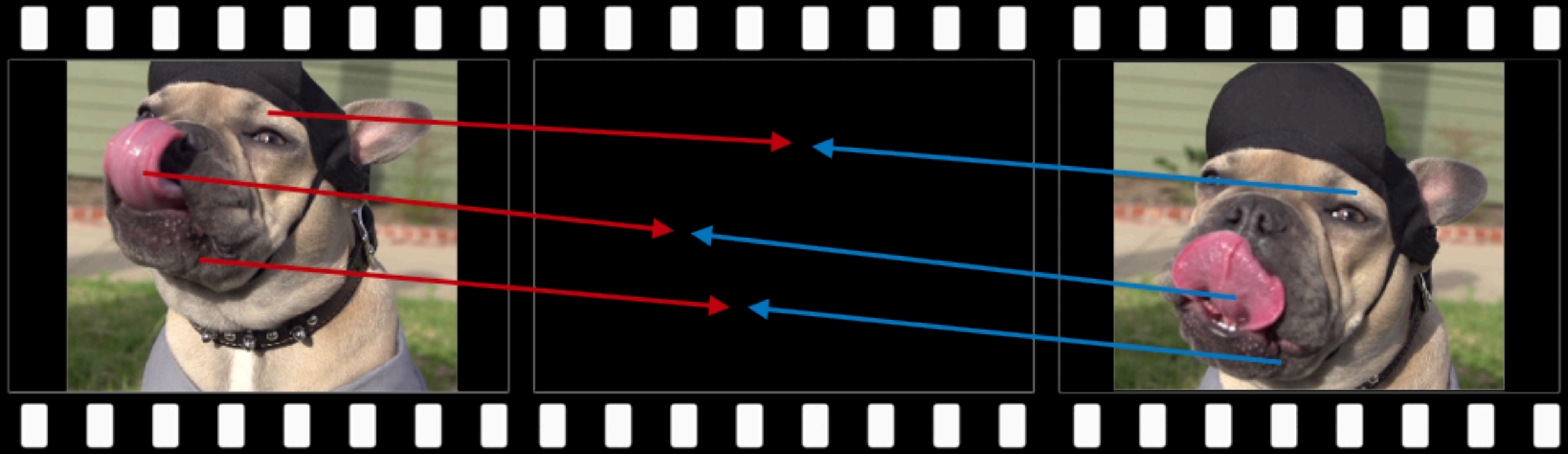


Video Frame Synthesis

- Challenge
  1. Complex motion (camera motion & scene motion)
  2. High-res images (1280 * 720)

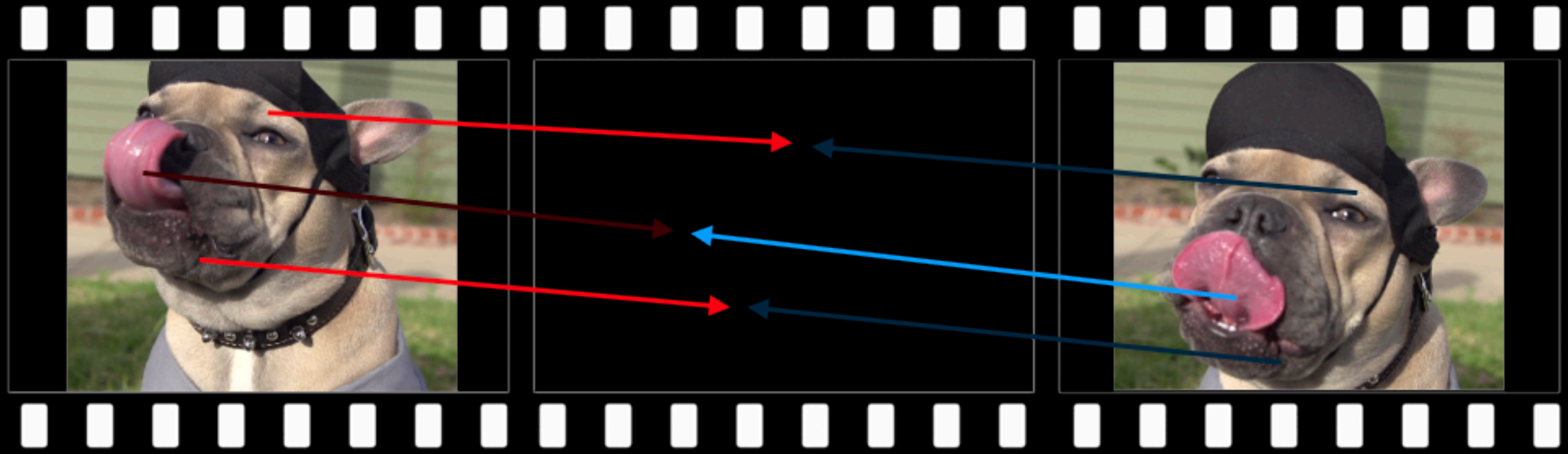


Voxel Flow



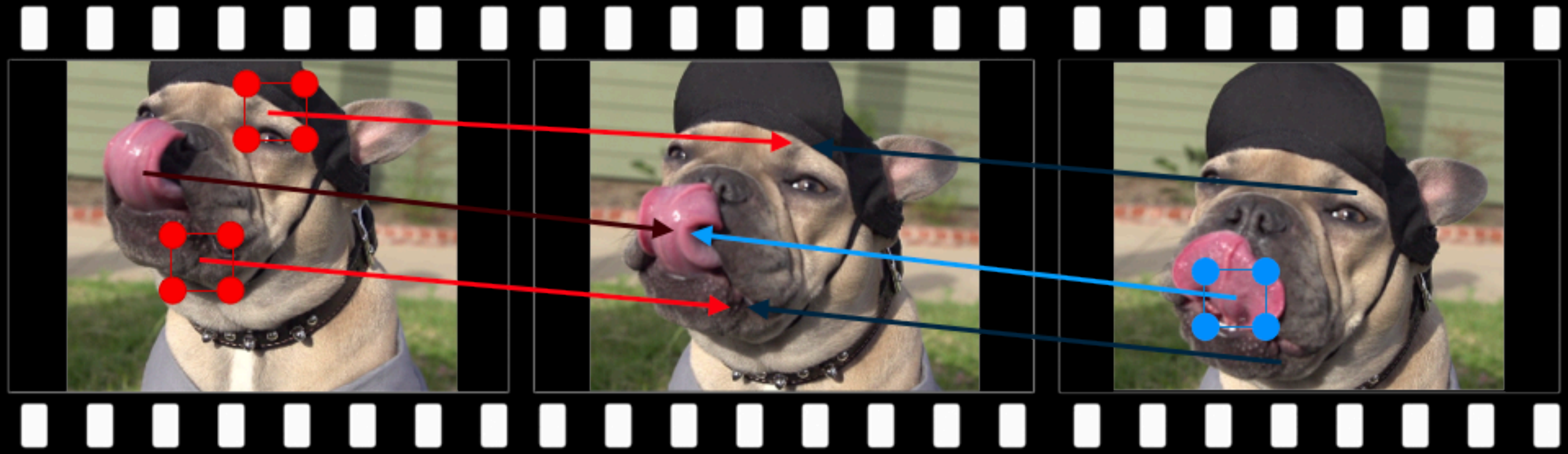
symmetric bi-directional flows

Voxel Flow



selection mask between frames

Voxel Flow

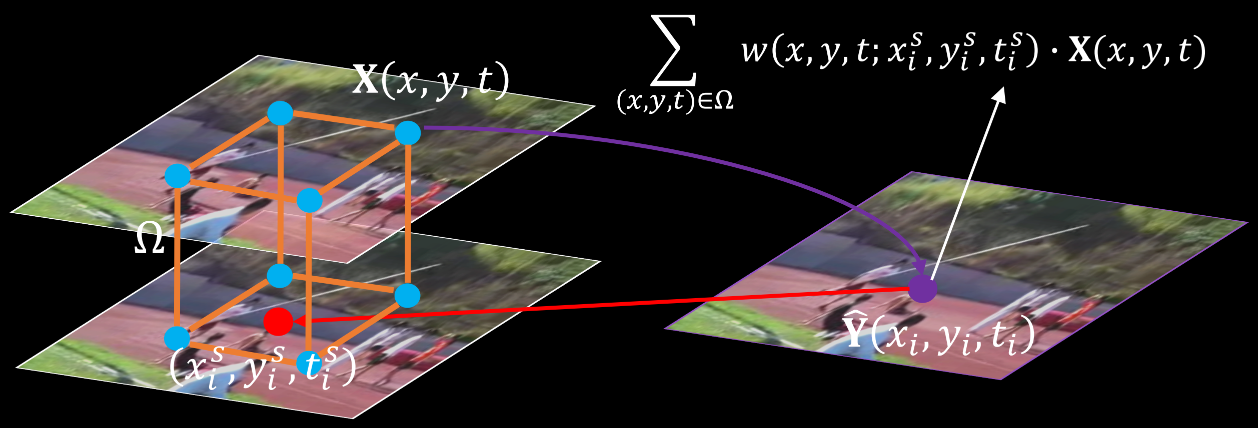


differentiable bilinear sampling

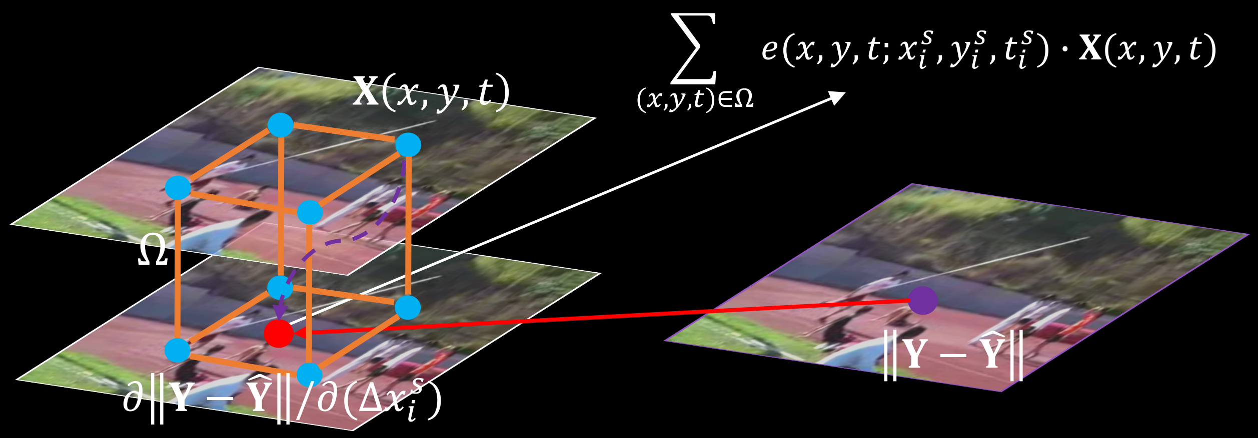
Deep Voxel Flow

- Mechanism

Differentiable spatio-temporal sampling



(a) Forward Pass

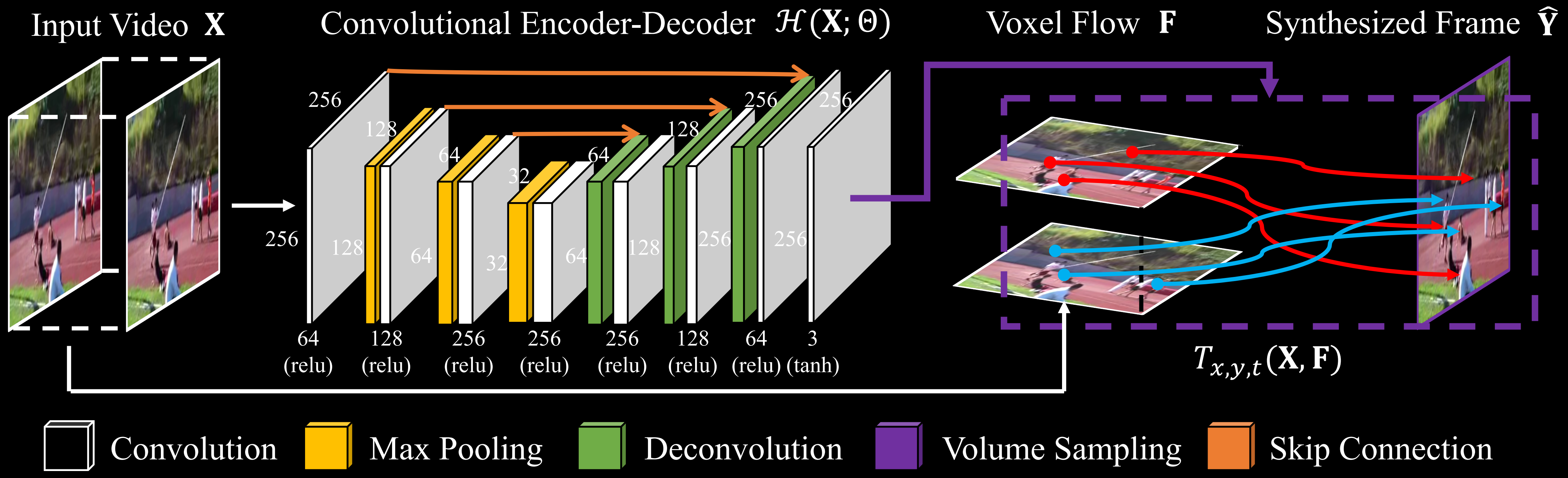


(b) Backward Pass

Deep Voxel Flow

- Motivation

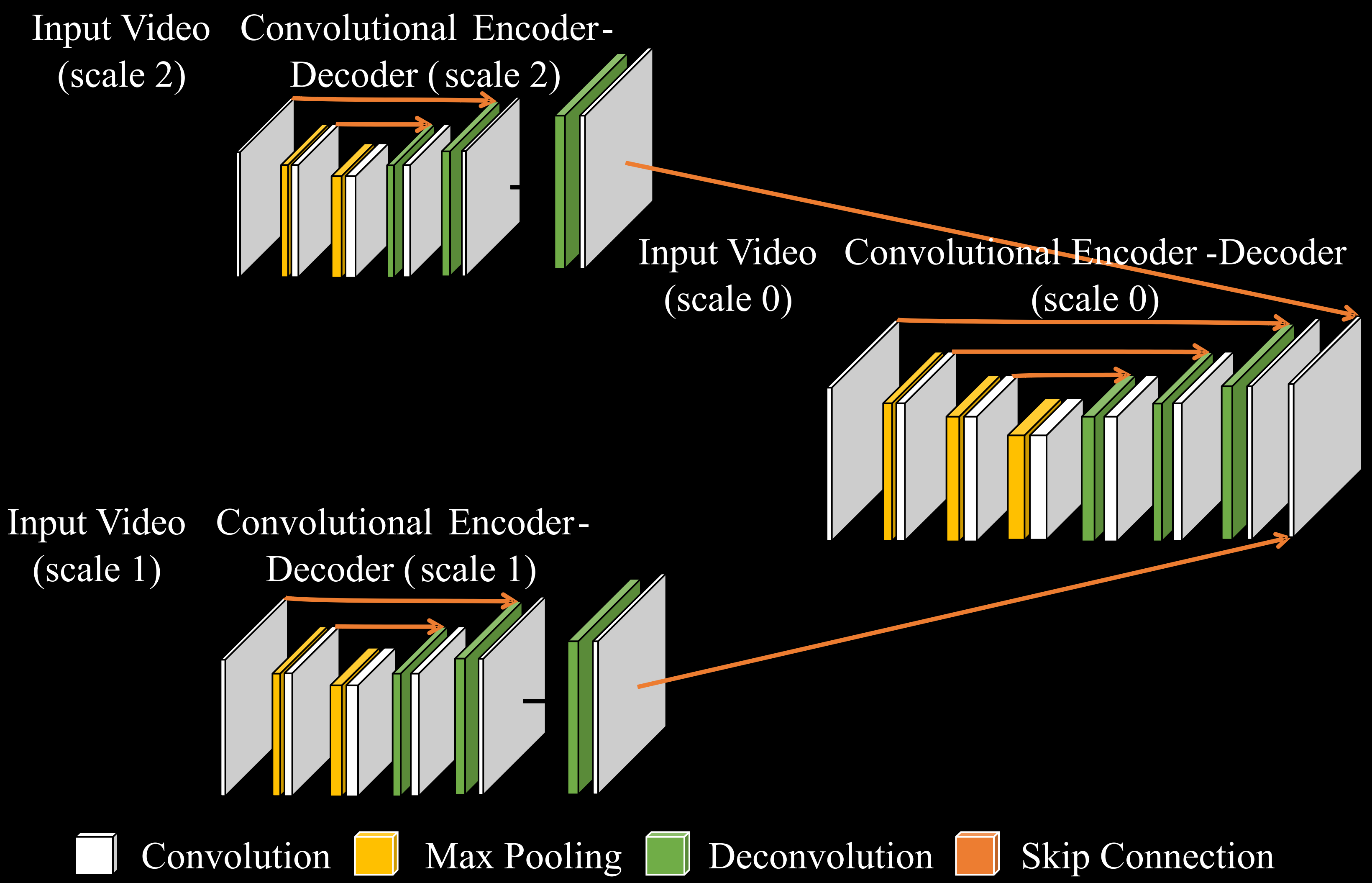
Combining the strength of flow-based and NN-based methods



Multi-scale Deep Voxel Flow

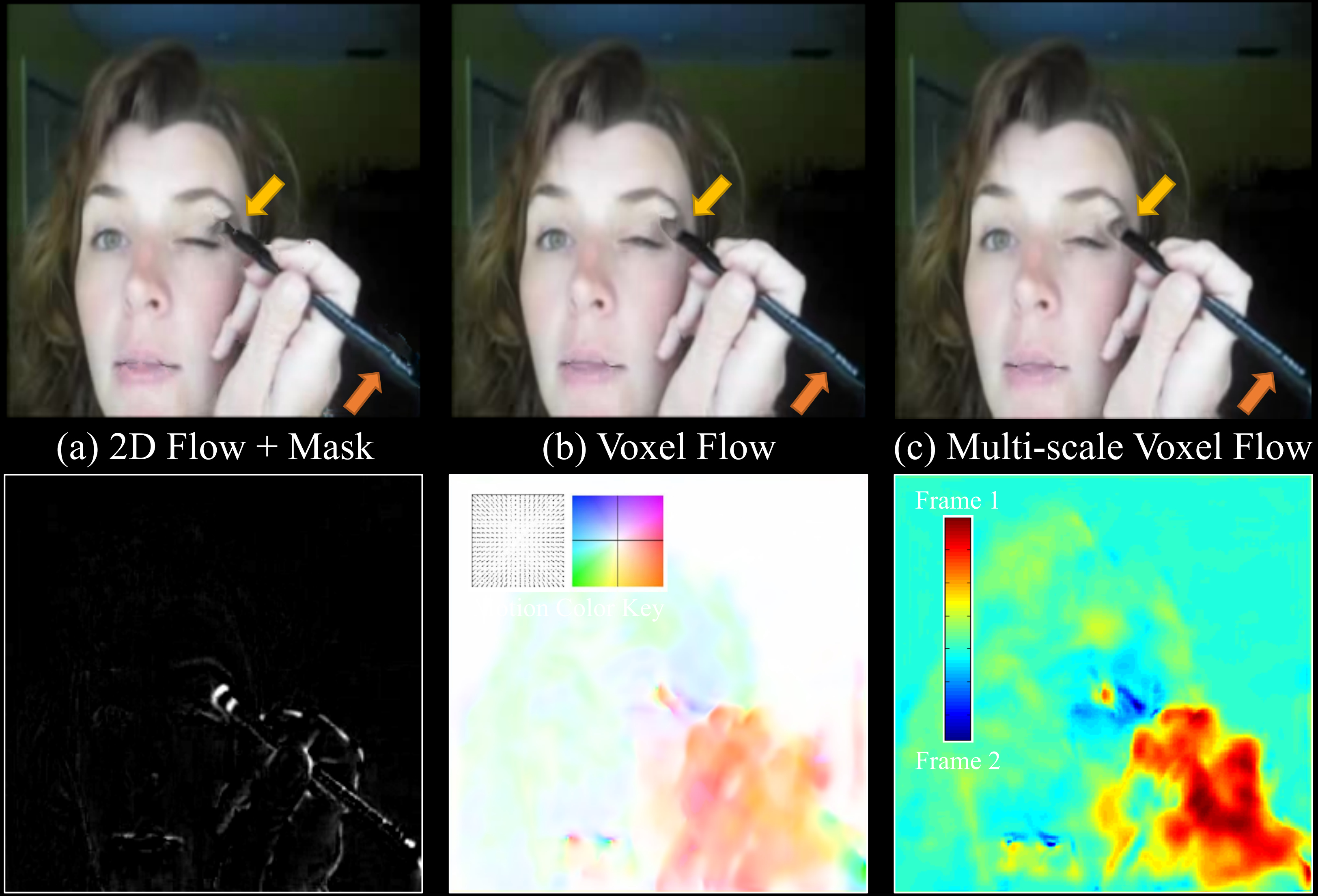
- Pipeline

Handle large motion



Multi-scale Voxel Flow

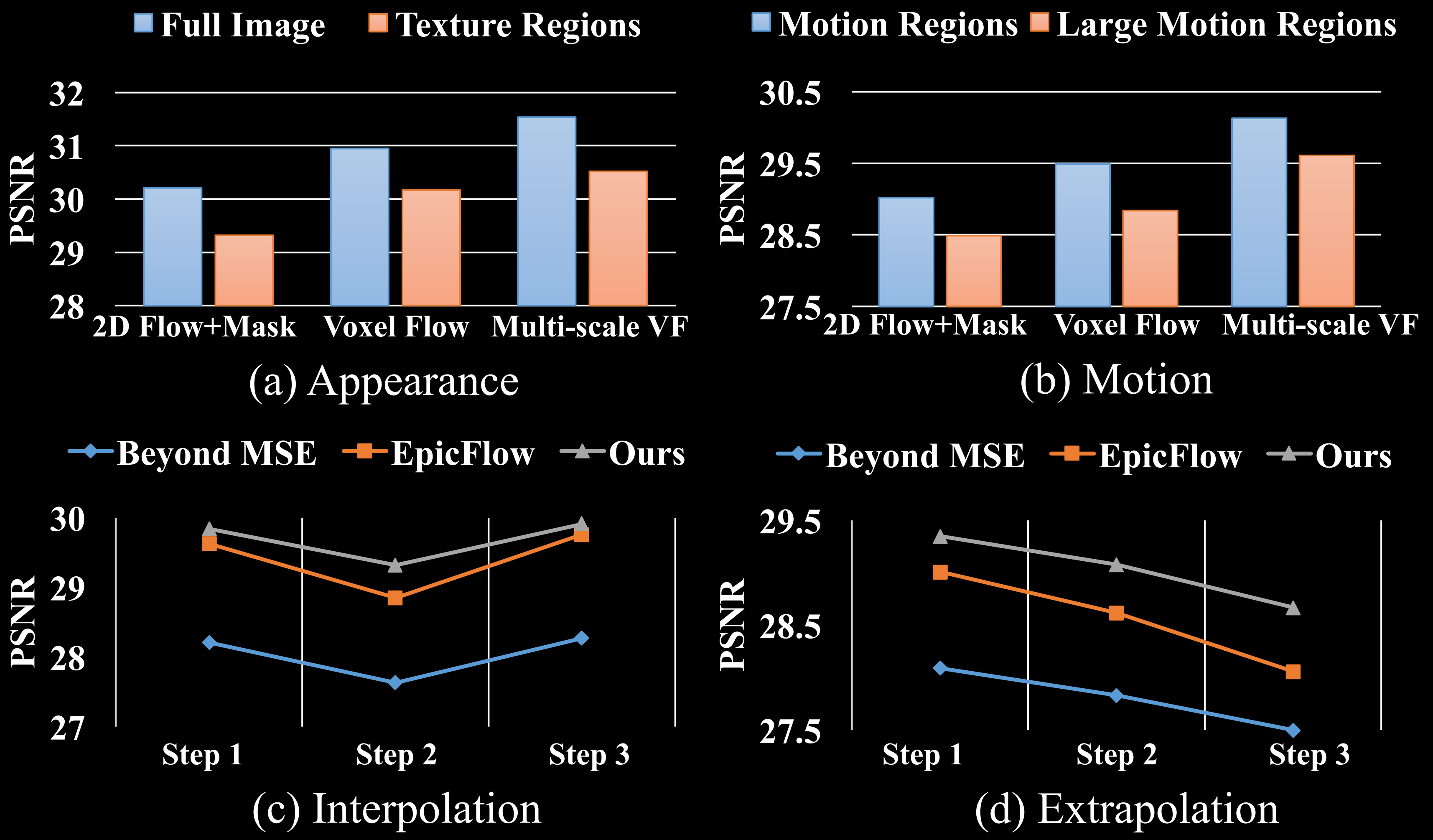
- Advantages



(a) 2D Flow + Mask (b) Voxel Flow (c) Multi-scale Voxel Flow
(d) Difference Image (e) Projected Motion Field (f) Projected Selection Mask

Multi-scale Voxel Flow

- Ablation Study



Comparisons

- UCF-101

Comparisons

- UCF-101

Comparisons

- KITTI

Comparisons

- KITTI

Feature Learning

- Self-supervised Learning

Method	EPE
LD Flow [3]	12.4
FlowNet [5]	9.1
EpicFlow [22]	3.8
Ours (w/o ft.)	14.6
Ours	**9.5**

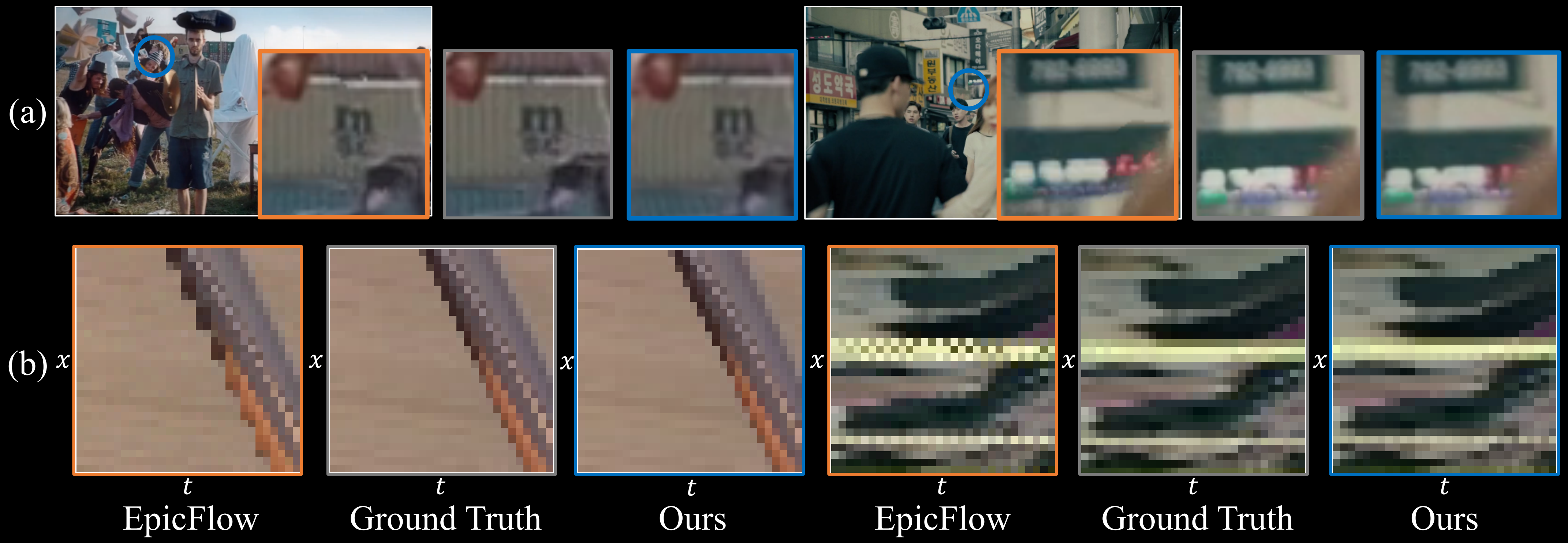
Flow estimation

Method	Acc.
Random	39.1
Unsup. Video [30]	43.8
ImageNet [14]	63.3
Ours (w/o ft.)	48.7
Ours	**52.4**

Action Recognition

Real-life Applications

- Spatio-temporal Coherence



Real-life Applications

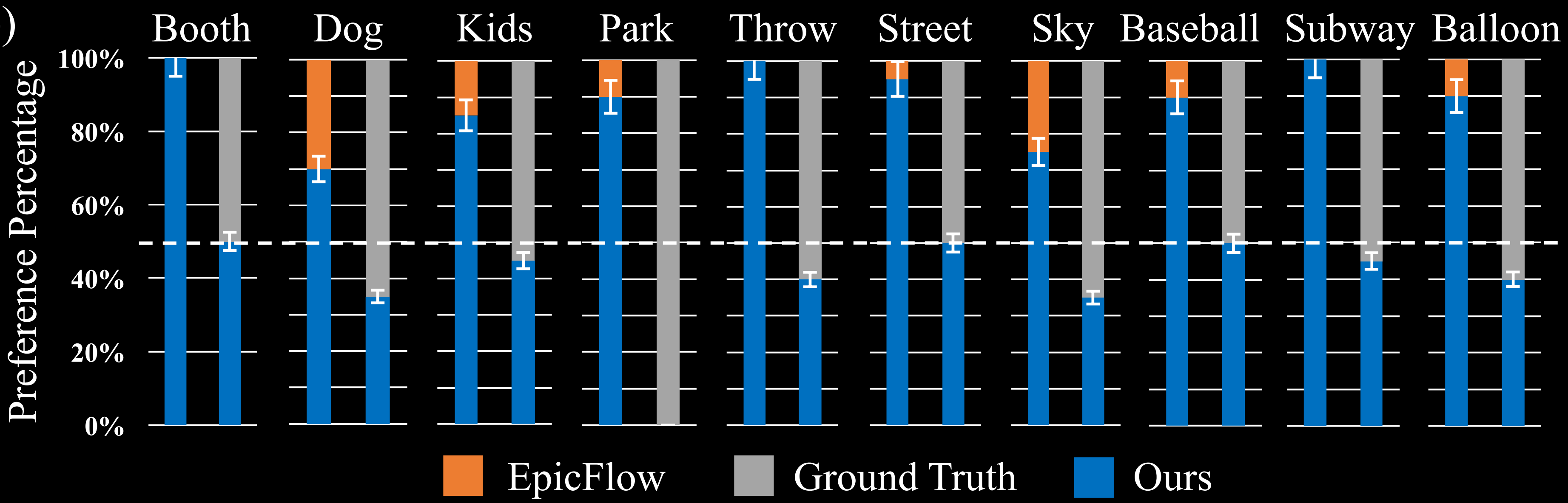
- User Study

(a)

Diagonal-split Comparison

Method 1 \ Method 2

(b)



Real-life Applications

Video Frame Synthesis using Deep Voxel Flow

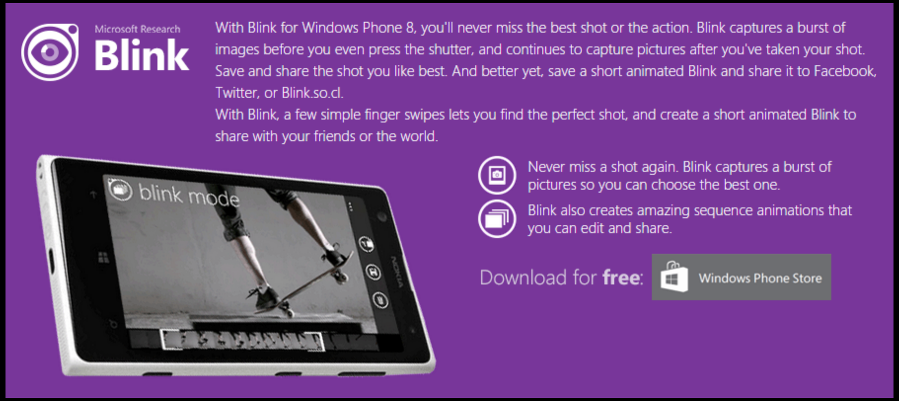
Ziwei Liu[1], Raymond Yeh[2], Xiaoou Tang[1], Yiming Liu[3], Aseem Agarwala[3]

[1]The Chinese University of Hong Kong
[2]University of Illinois at Urbana-Champaign
[3]Google

Conclusions & Future Work

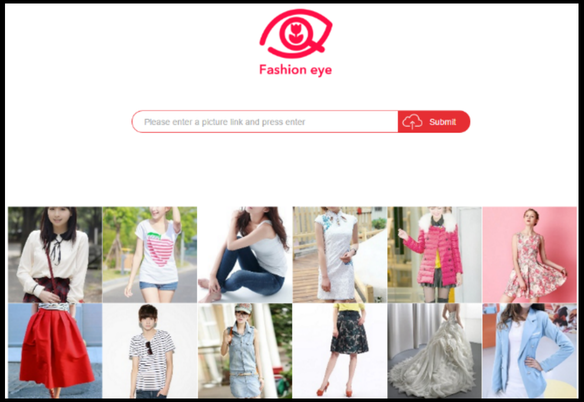
- In-the-Wild Handling: deformable objects, complex scenes
- Heter. Supervisions: identity, attribute, landmark, self-sup
- Structural Deep Learning: semantic, geometry, spatio-temporal

Product Transfer



Microsoft Blink

Google Clips



SenseTime FashionEye

Collaborators


Xiaoxiao Li


Sijie Yan


Shi Qiu


Ping Luo


Chen Change Loy


Xiaogang Wang


Xiaoou Tang

Thanks!

Science is what we understand well enough to explain to a computer.
Art is everything else we do.

Homepage: https://liuziwei7.github.io/