Deep Learning Human-centric Representation in the Wild

Ziwei Liu

Multimedia Lab, The Chinese University of Hong Kong

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

Face Attributes Recognition

- Problem

Arched Eyebrows?

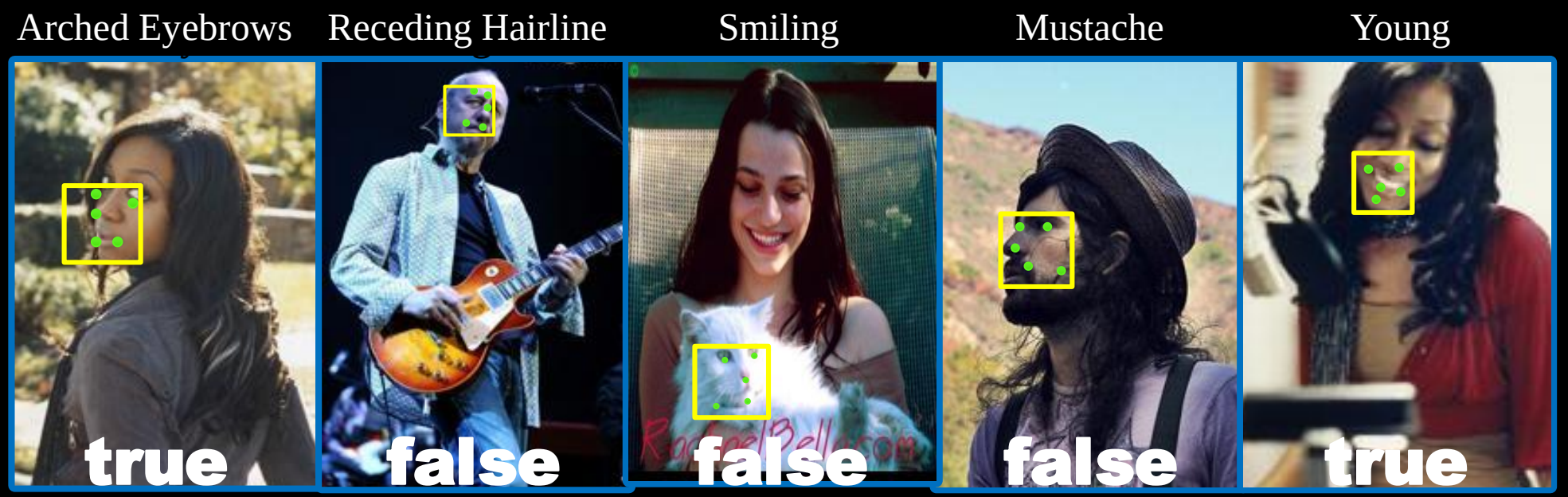
Big Eyes?

Receding Hairline?

Mustache?

Face Attributes Recognition

- Challenges



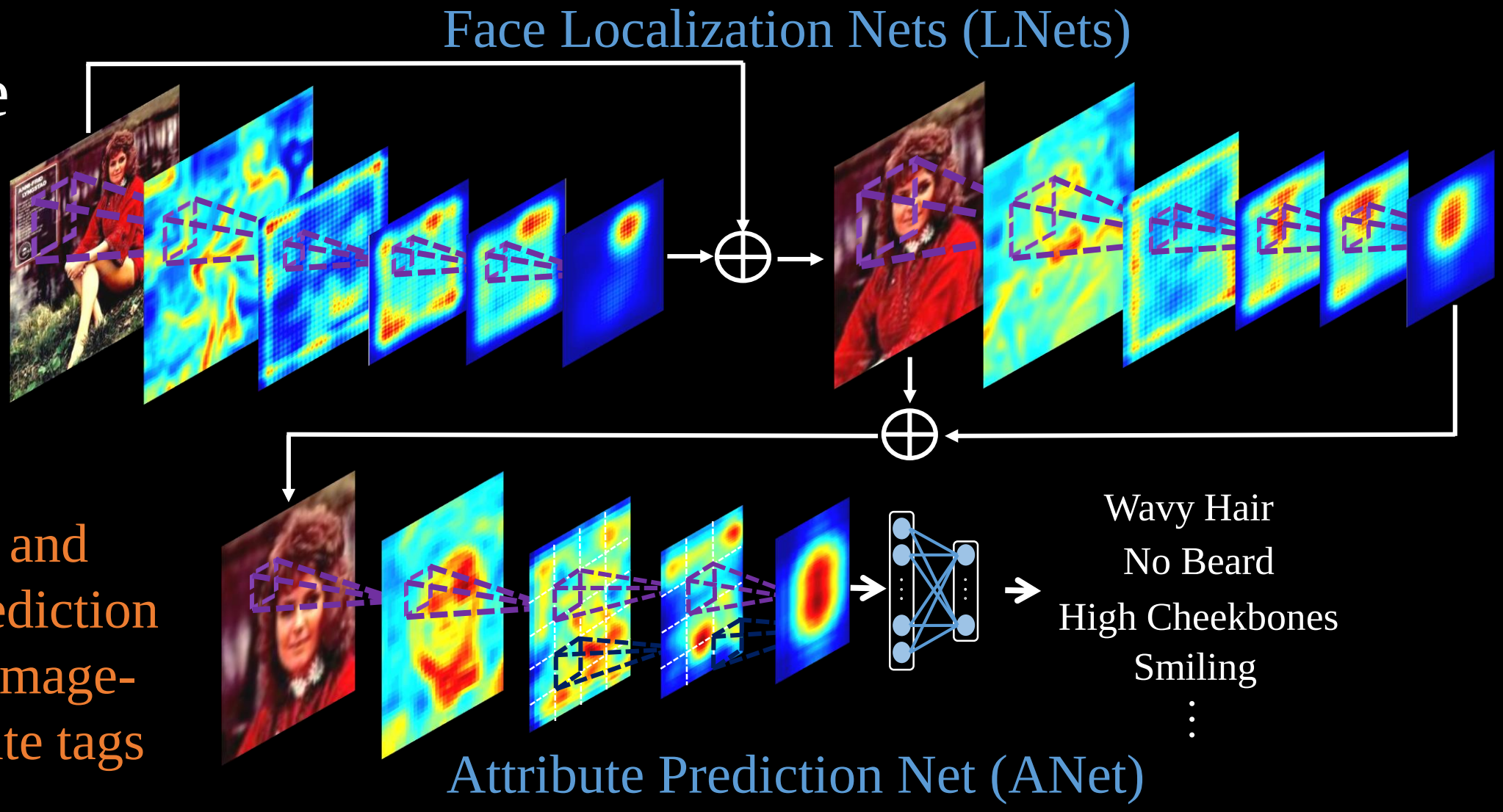
HOG (landmarks) + SVM

Face Attributes Recognition

- Motivation

Face Attributes Recognition

- Pipeline



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

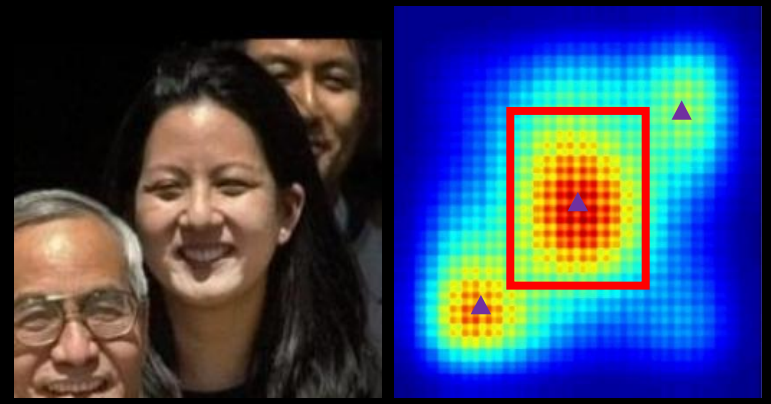
Face Attributes Recognition

- Technical Details

  1. Proposal-based face localization

  2. Multiple faces pruning

  3. Efficient feed-forward for locally-shared filters



$$\mathbf{h}^{(4)} = \arg\max_{\Omega_{(i,j)}} \max_{\forall(p,q)\in\Omega_{(i,j)}} \{\mathbf{h}^{(4)}_{inter}(\Omega_{(i,j)})\},$$

Face Attributes Recognition

- Advantages

1. An end-to-end system

Integrate face localization and face attributes recognition into a seamless framework

2. Efficient arbitrary-sized evaluation

Devise efficient feed-forward techniques using globally convolution

Face Attributes Recognition

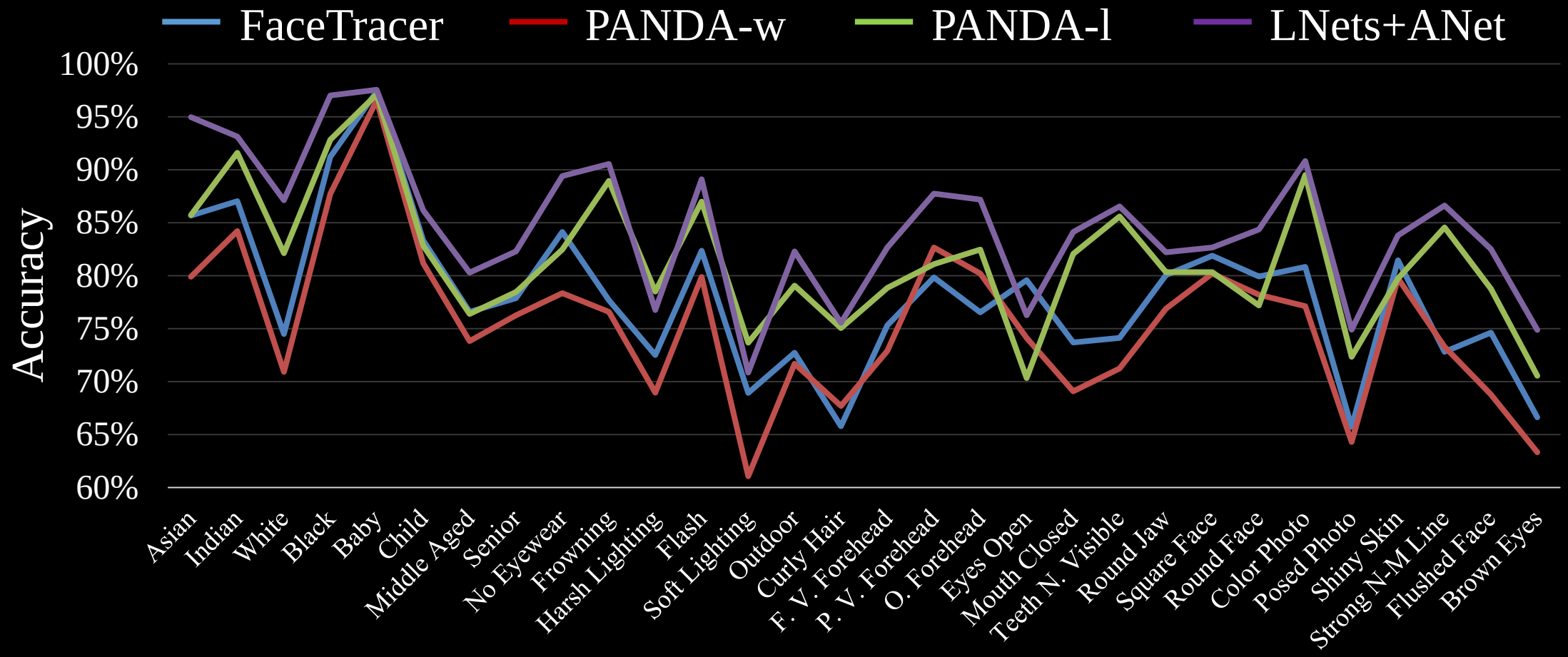
- Attribute recognition performance (40 attributes)

	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)

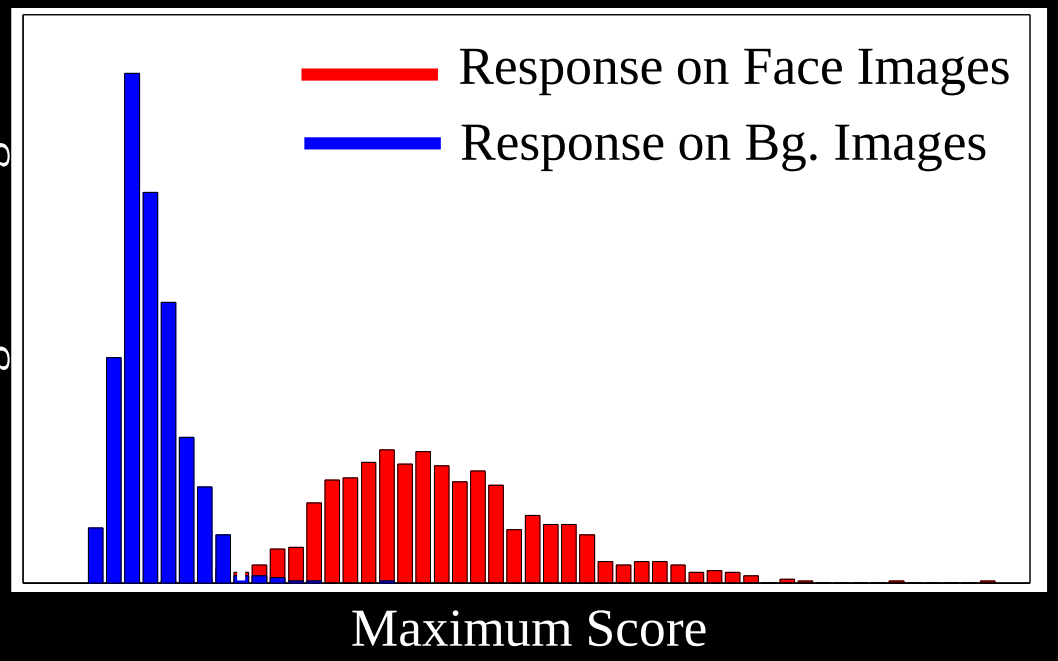
Face Attributes Recognition

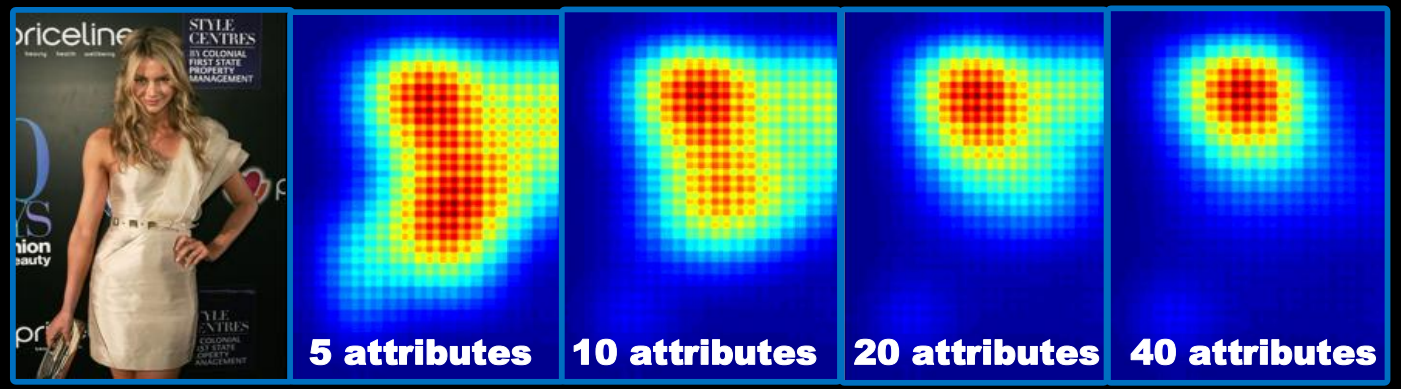
- Performance on unseen attributes (30 attributes)



Deep Face Representation

- Rich attributes tags enable accurate face localization





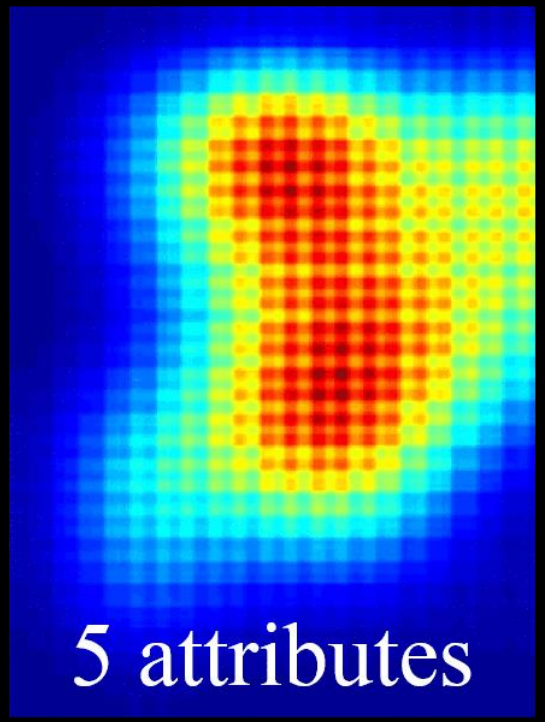
Response map with different numbers of attributes

Deep Face Representation

- Rich attributes tags enable accurate face localization

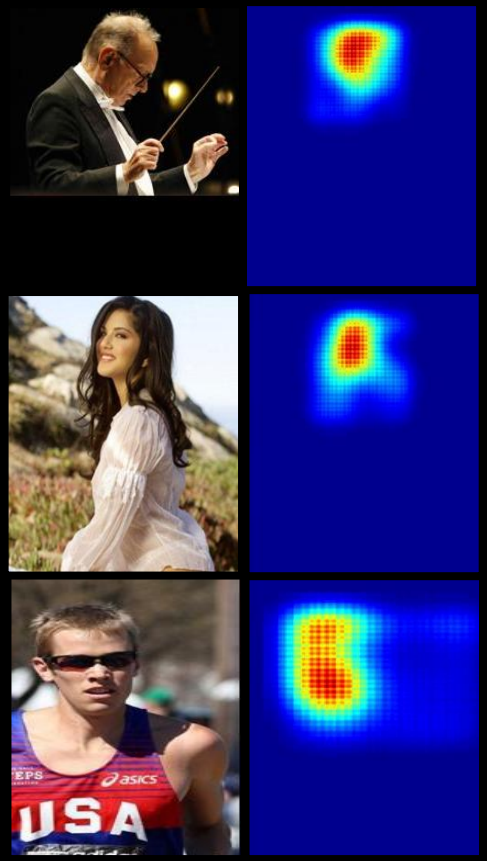
Original Image

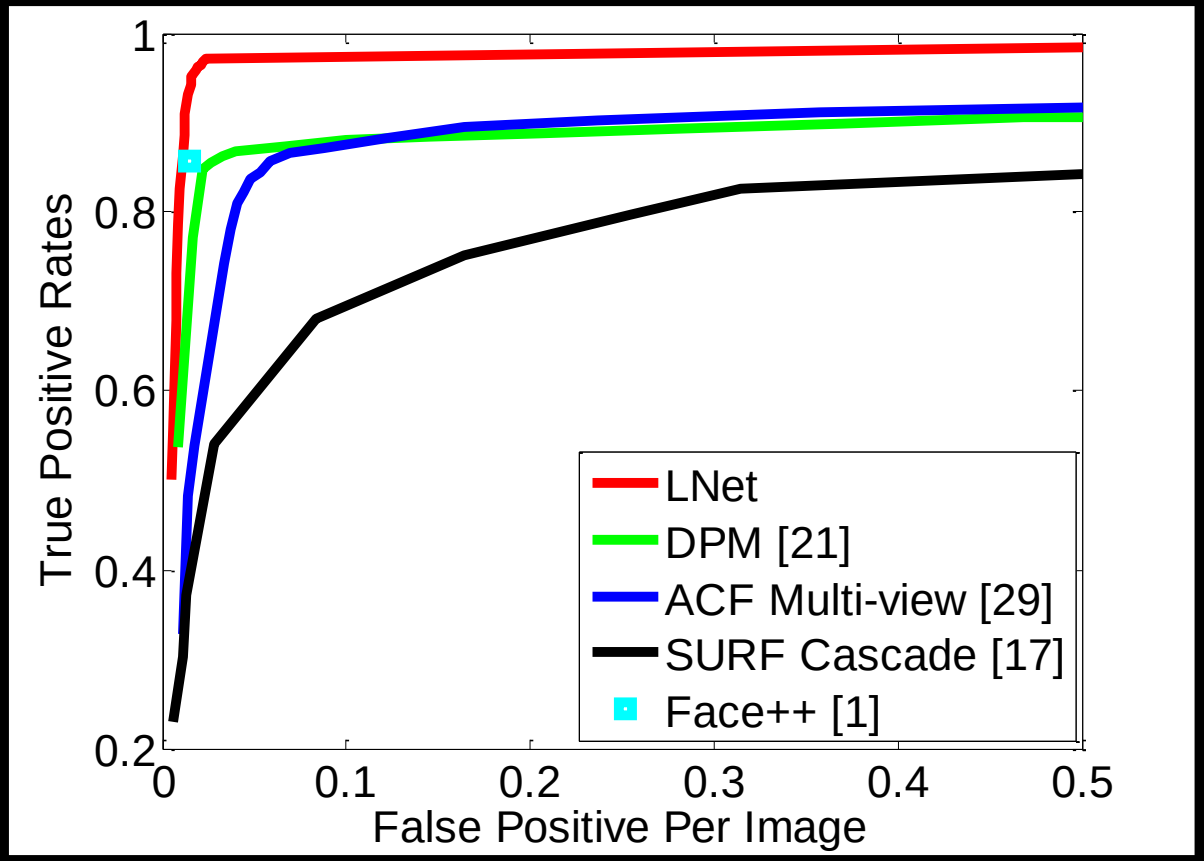
Response Map



Deep Face Representation

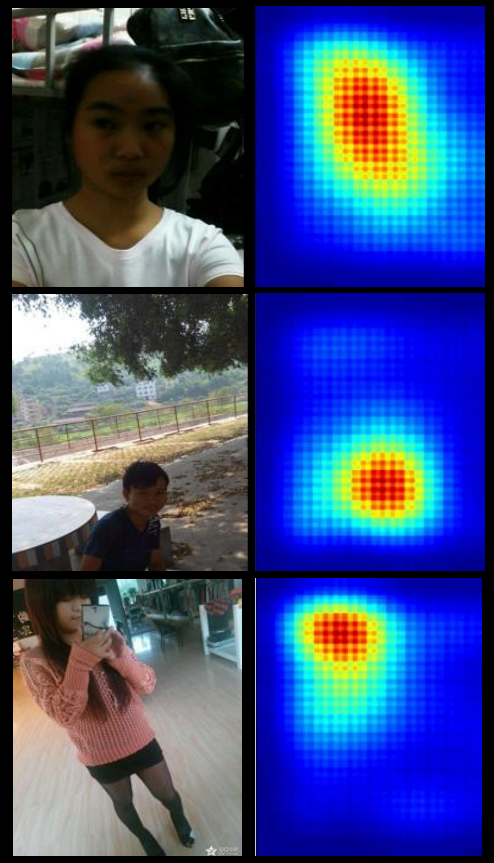
- Face localization performance on CelebFace

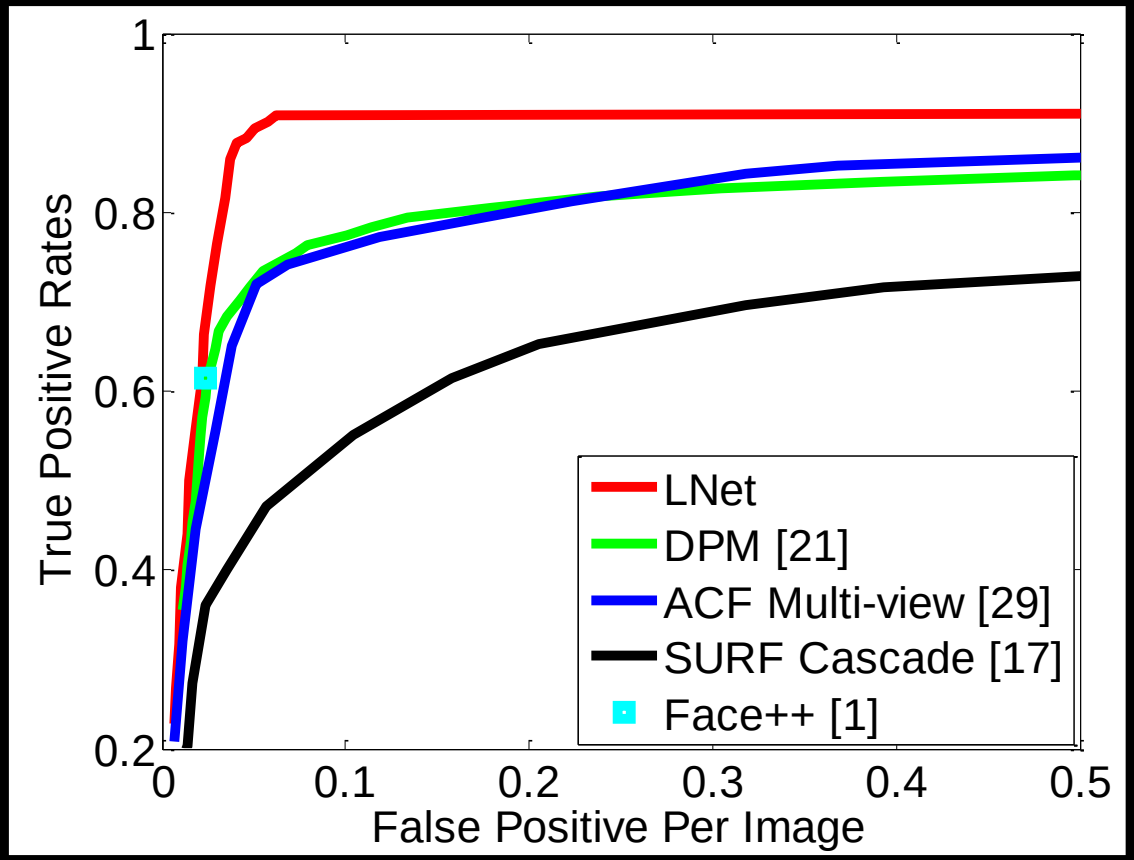




Deep Face Representation

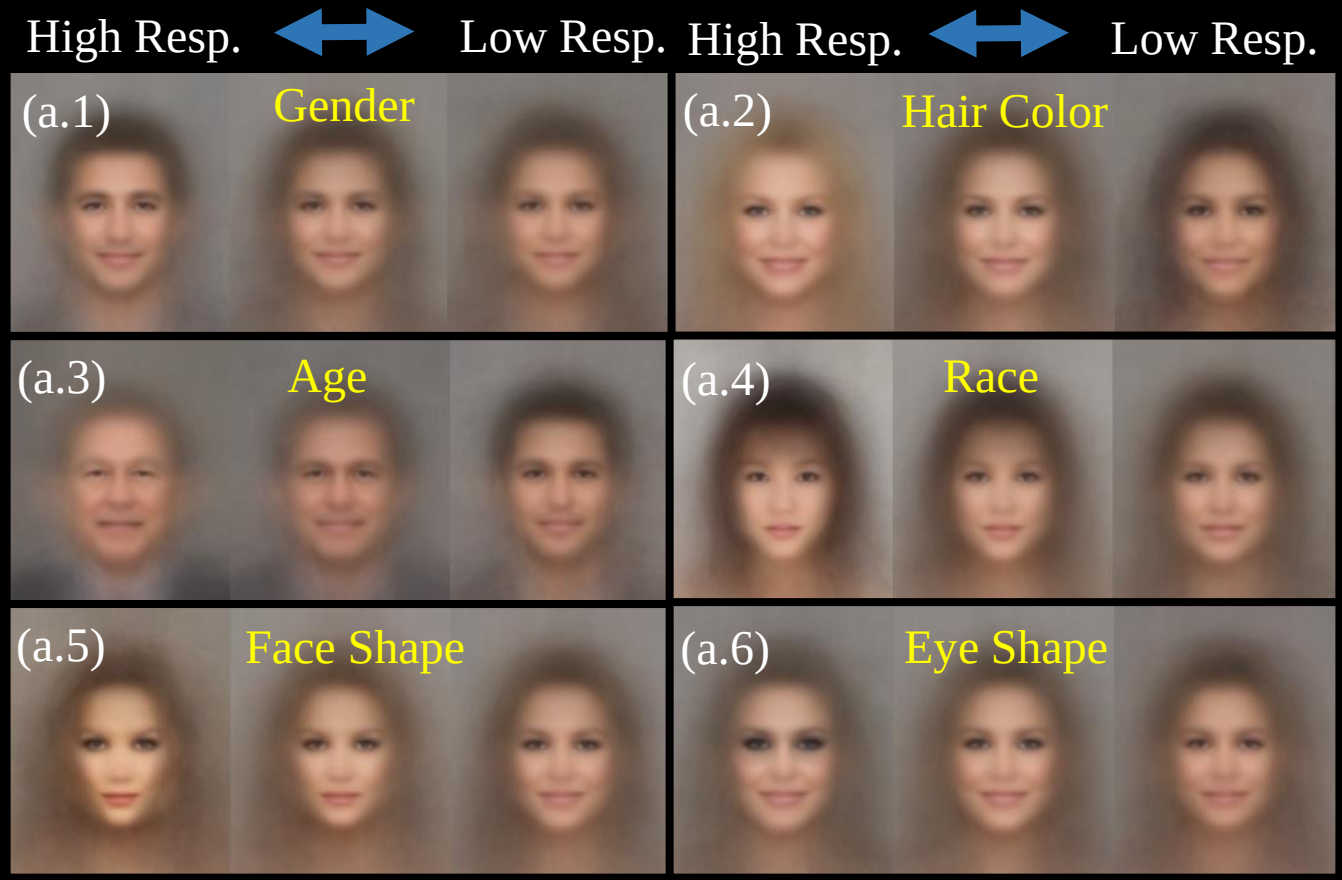
- Face localization performance on MobileFace





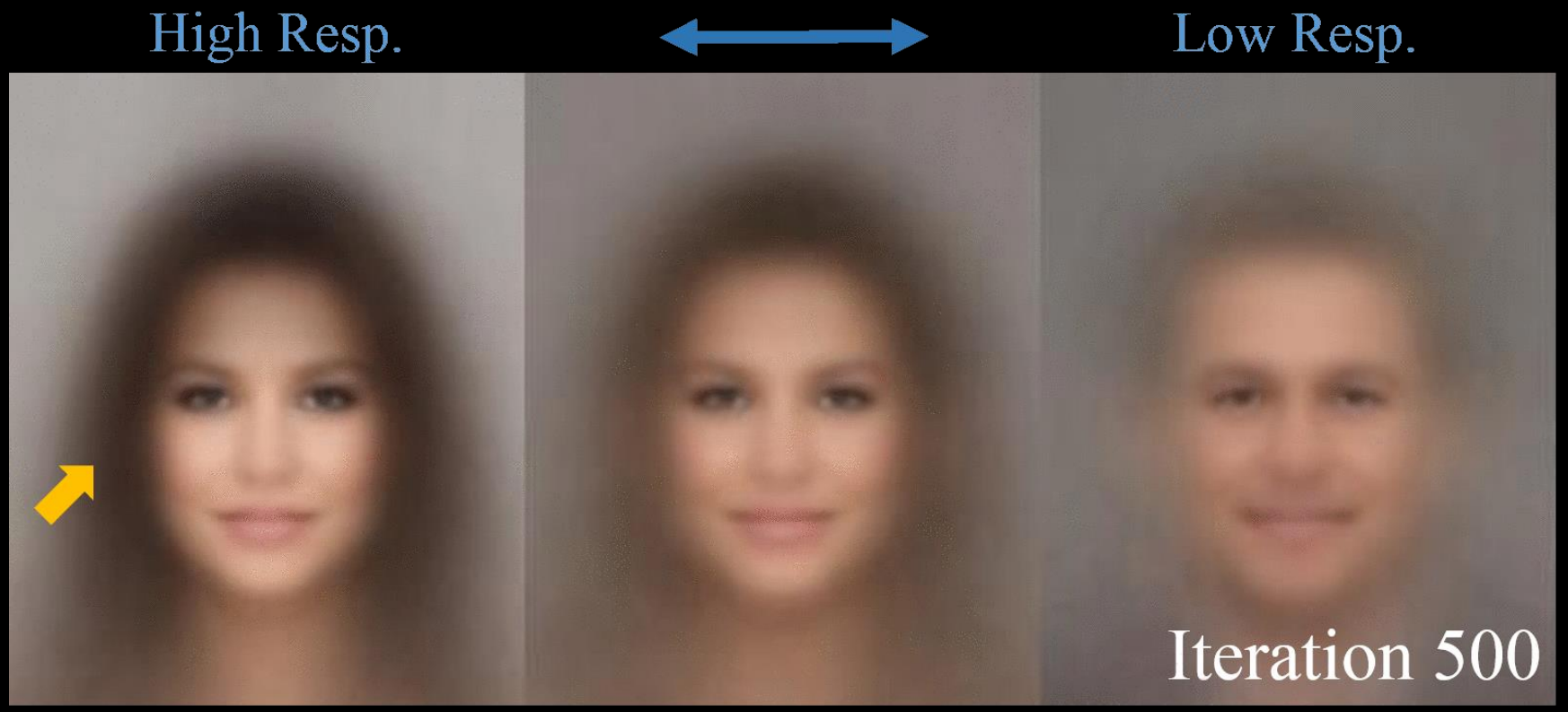
Deep Face Representation

- Pre-training with identities discovers semantic concepts



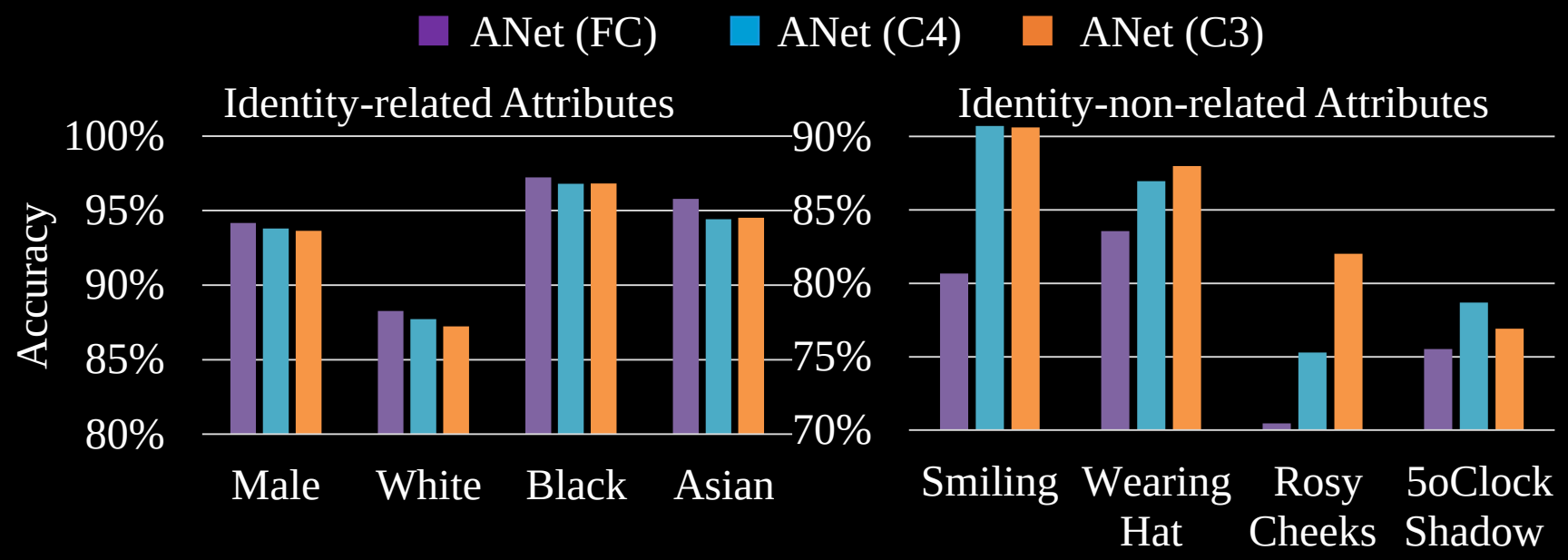
Deep Face Representation

- Pre-training with identities discovers semantic concepts



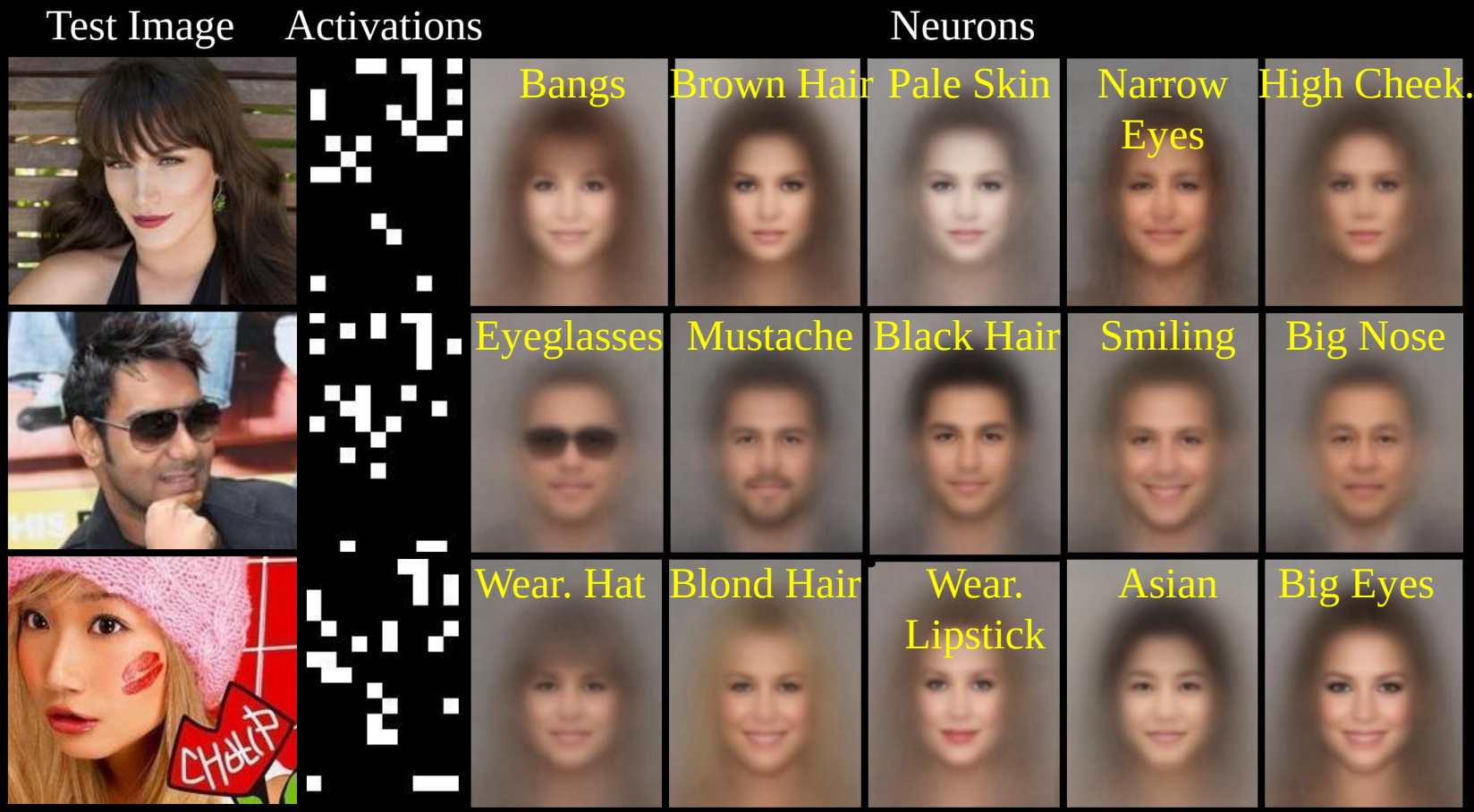
Deep Face Representation

- Pre-trained concepts have layer-wise distribution



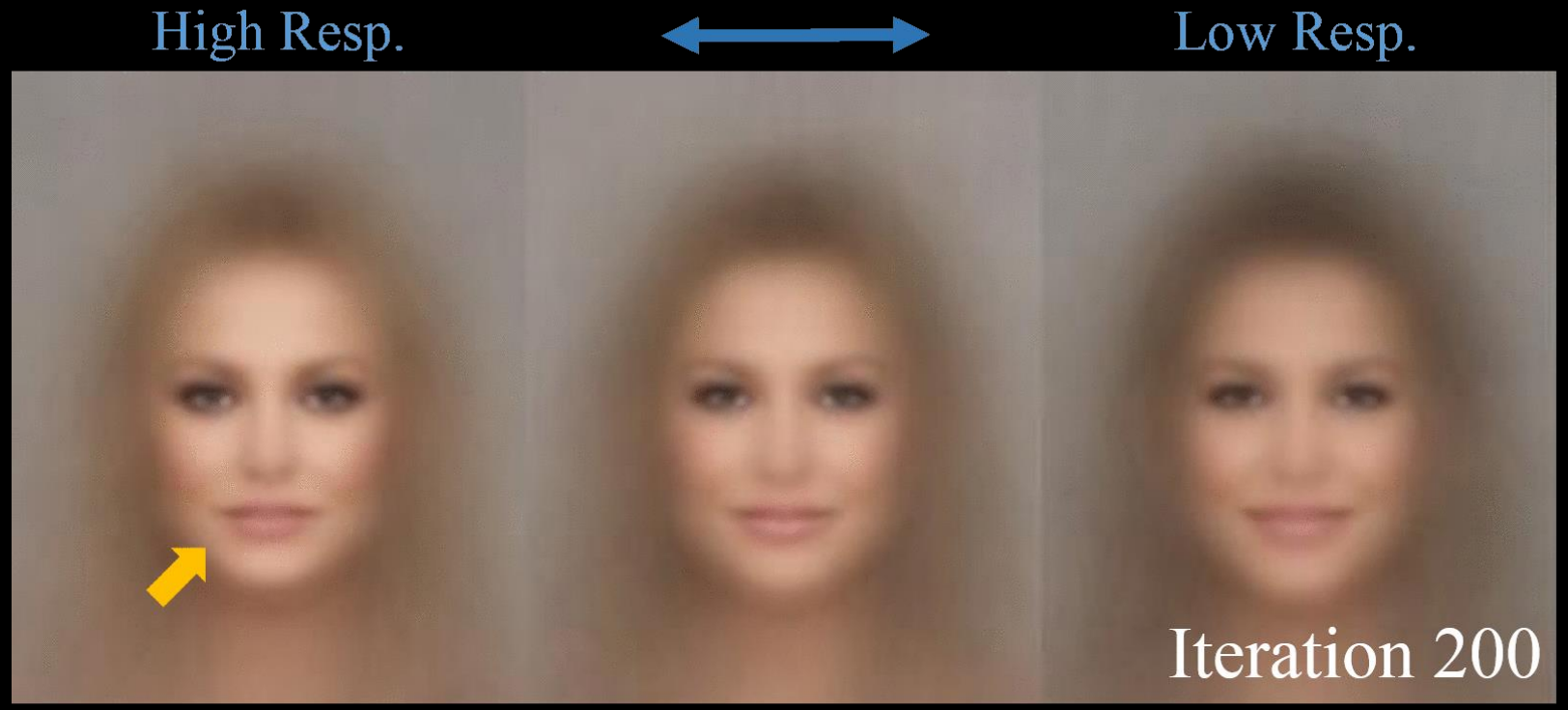
Deep Face Representation

- Fine-tuning with attributes expands semantic concepts



Deep Face Representation

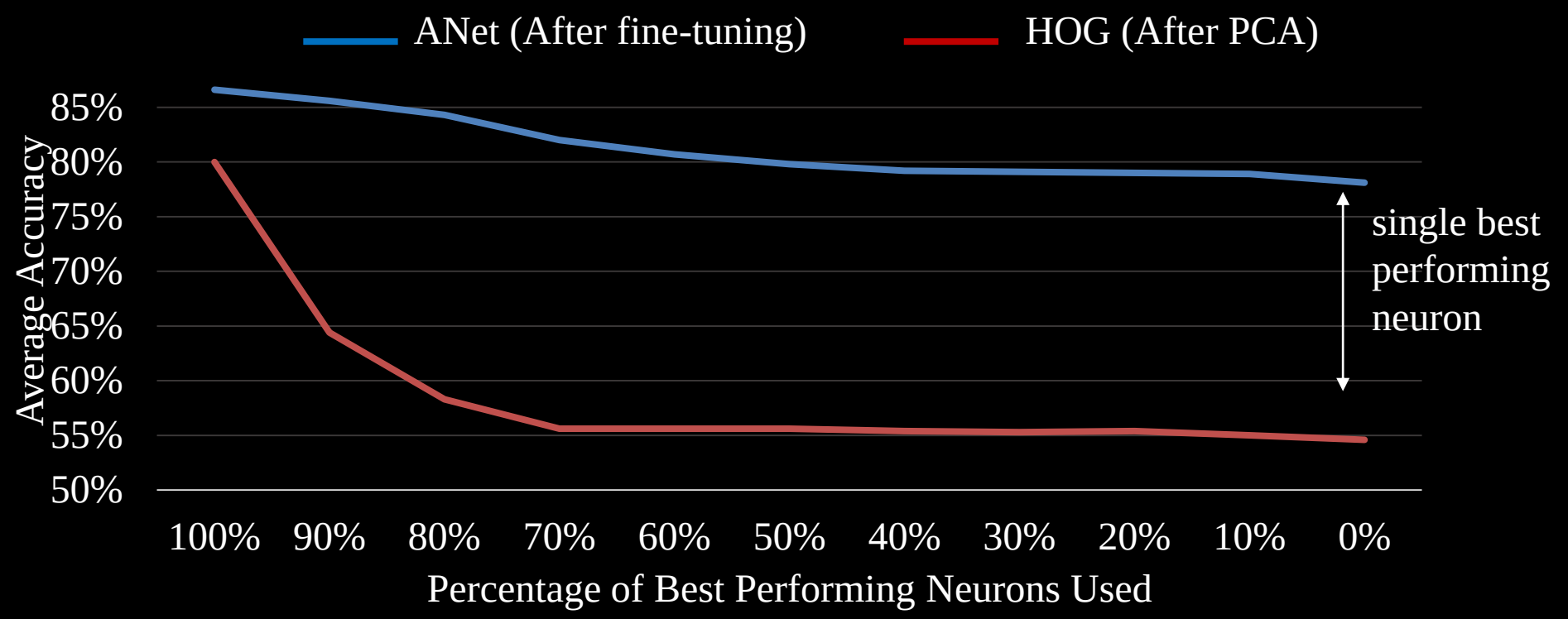
- Fine-tuning with attributes expands semantic concepts



Thick Lip (Neuron #152)

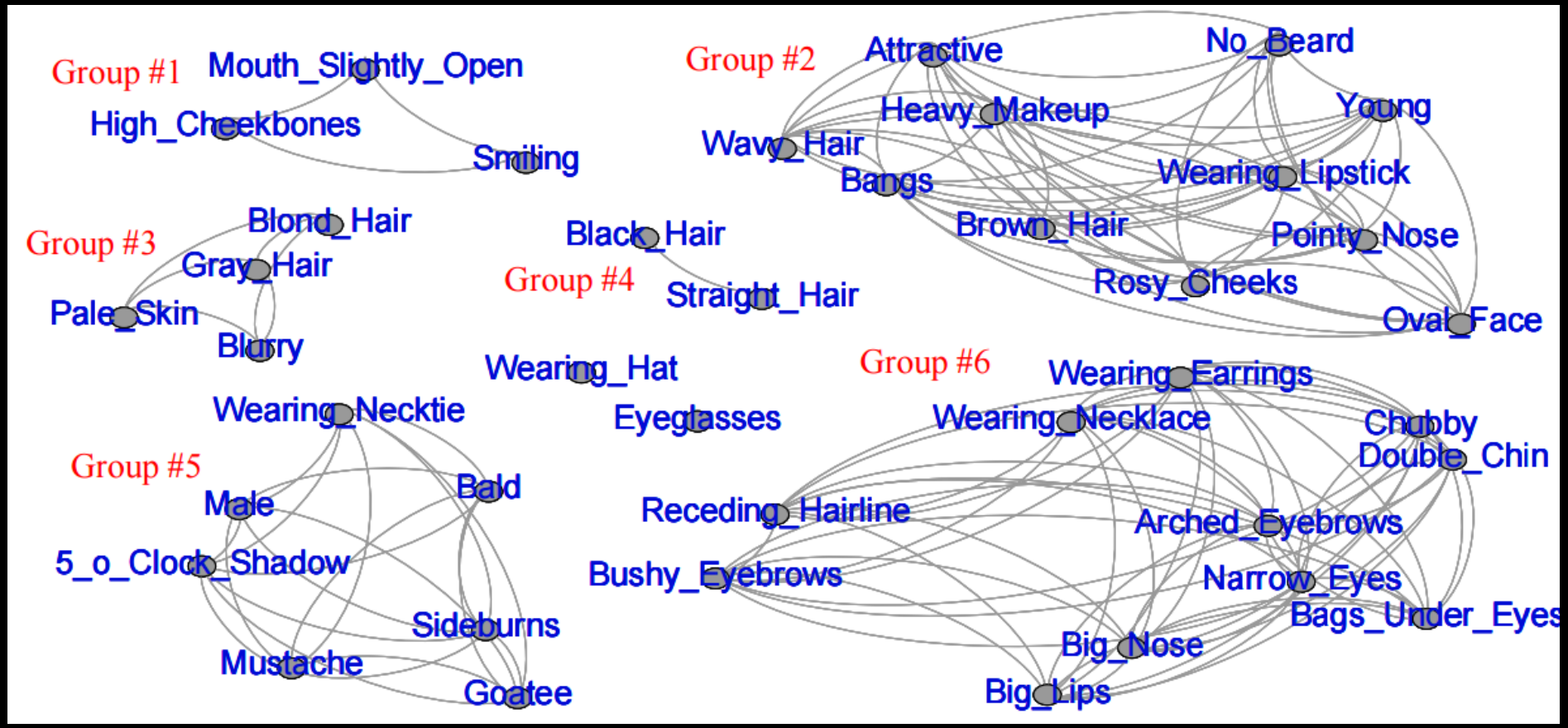
Deep Face Representation

- Fine-tuned neurons have strong attribute indication



Deep Face Representation

- Attributes groups are automatically discovered



Part II: Deep Fashion Understanding

Challenges

Face Variations

Cloth Variations

Overall Pipeline

Clothes Detection

Overall Pipeline

Clothes Detection

Clothes Alignment

Overall Pipeline

Clothes Recognition

Clothes Alignment

A set of fashion landmarks

Collars
Cuffs
Waistlines
Hemlines
…

(a.1)

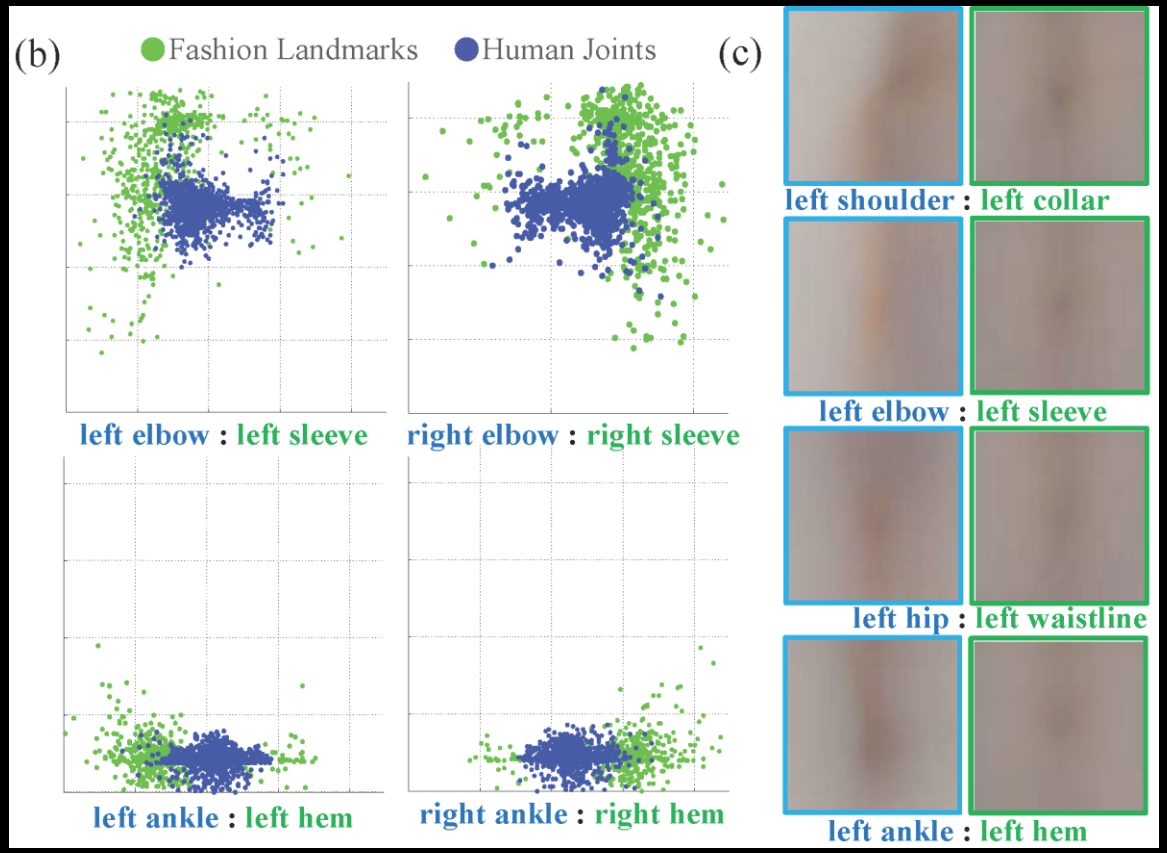
(a.2)

(a.3)

(a.4)

Clothes Alignment

More challenging than human pose estimation

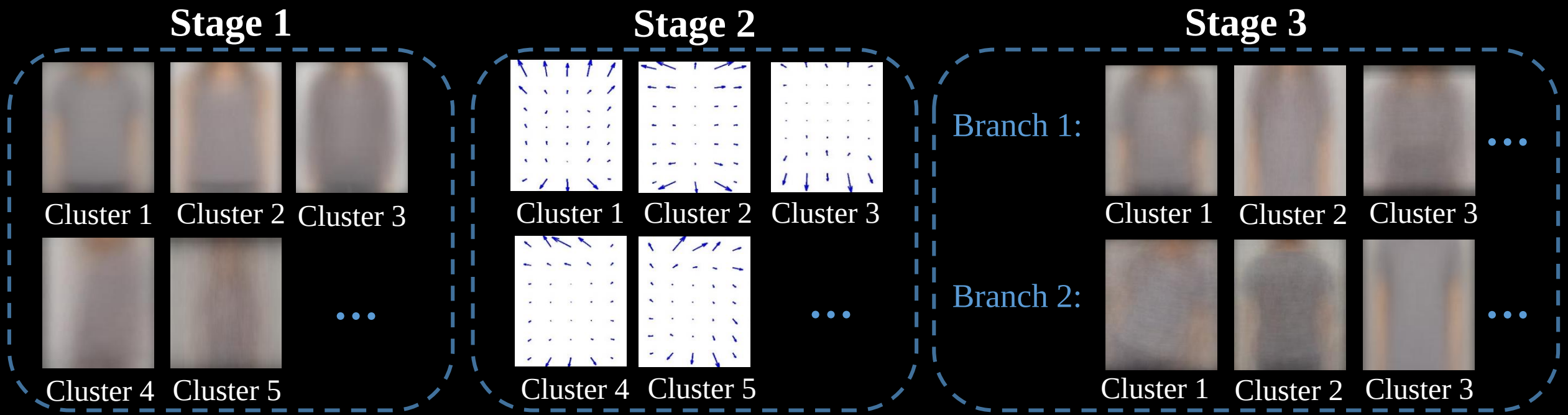


Geometry

Appearance

Clothes Alignment

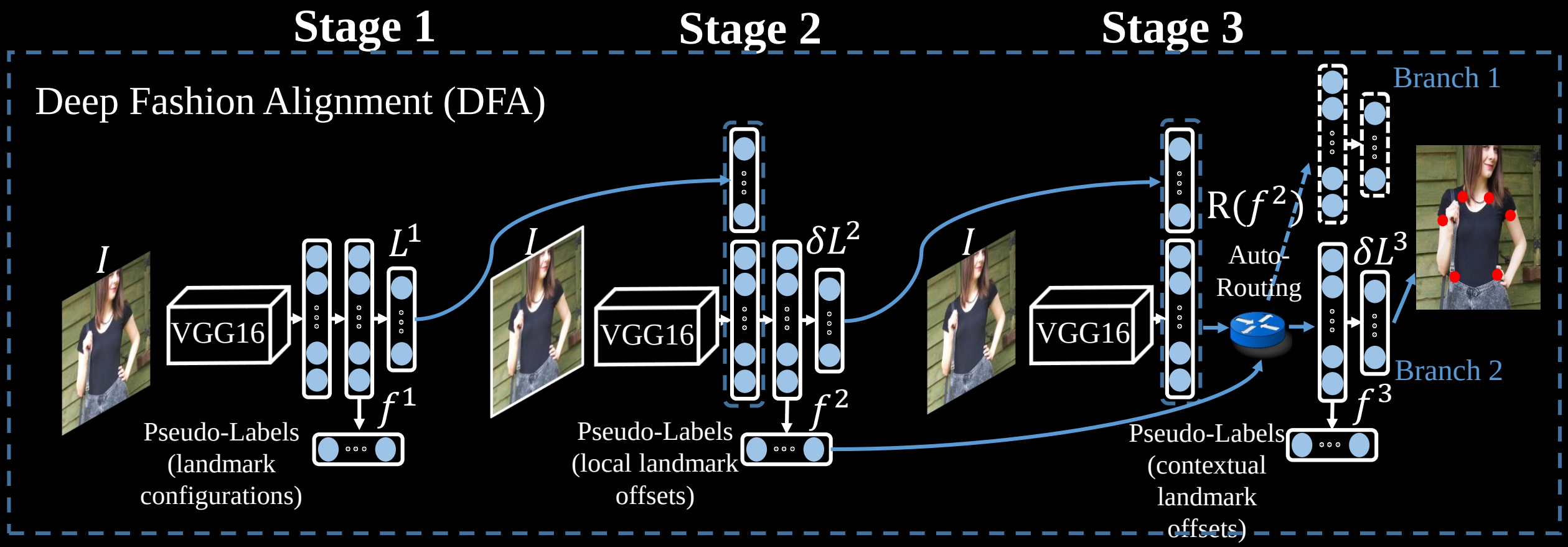
Reduce variations by pseudo-labels



Obtain codebook by k-means clustering in label space

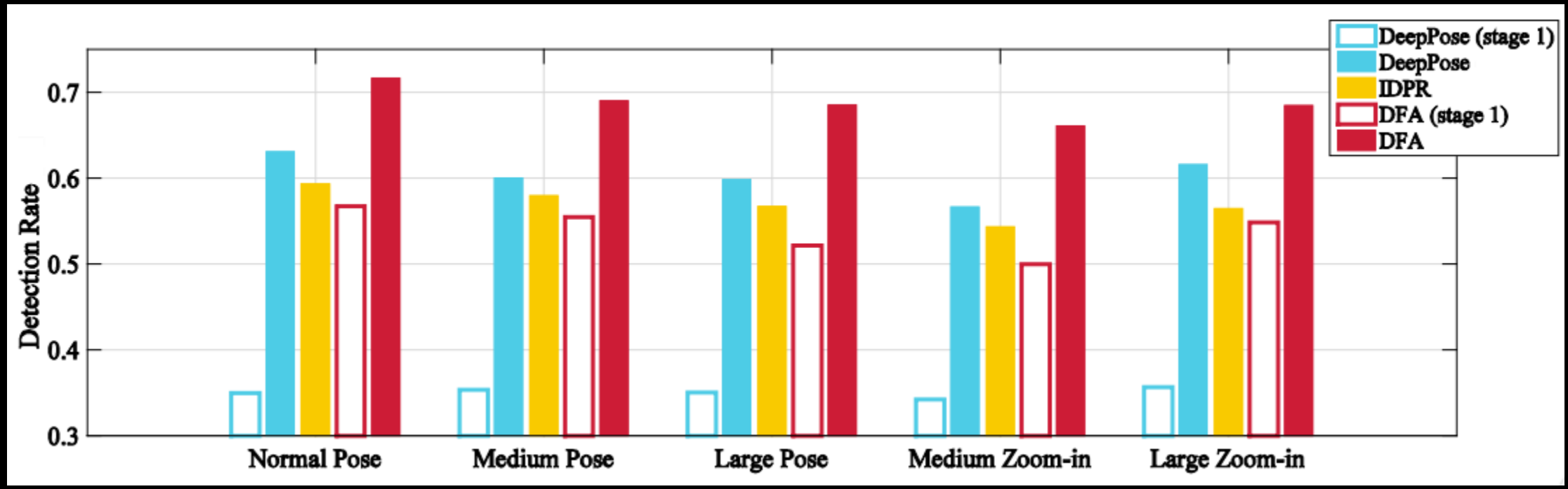
Clothes Alignment

Pipeline



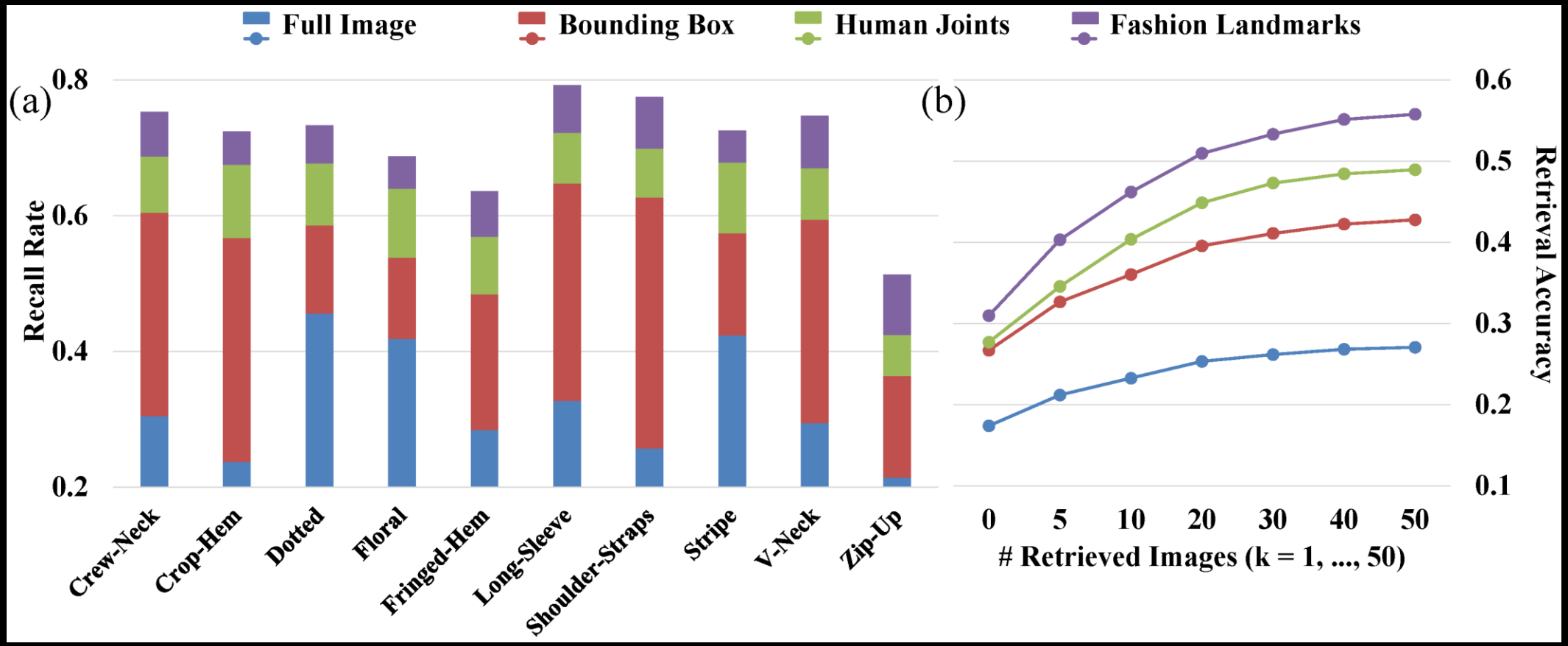
Clothes Alignment

Performance



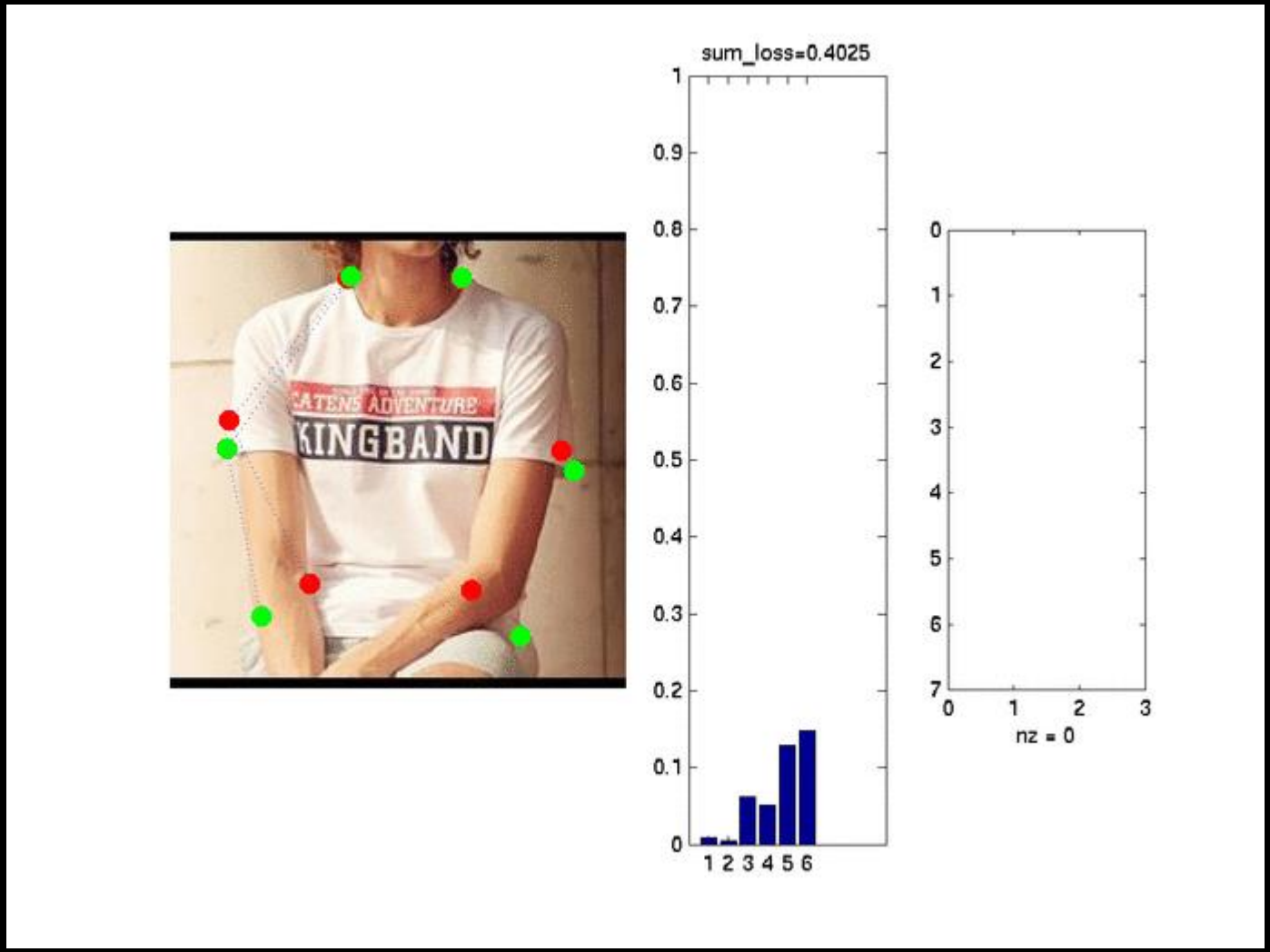
Clothes Alignment

More effective representation



Clothes Alignment

Demo



Clothes Recognition

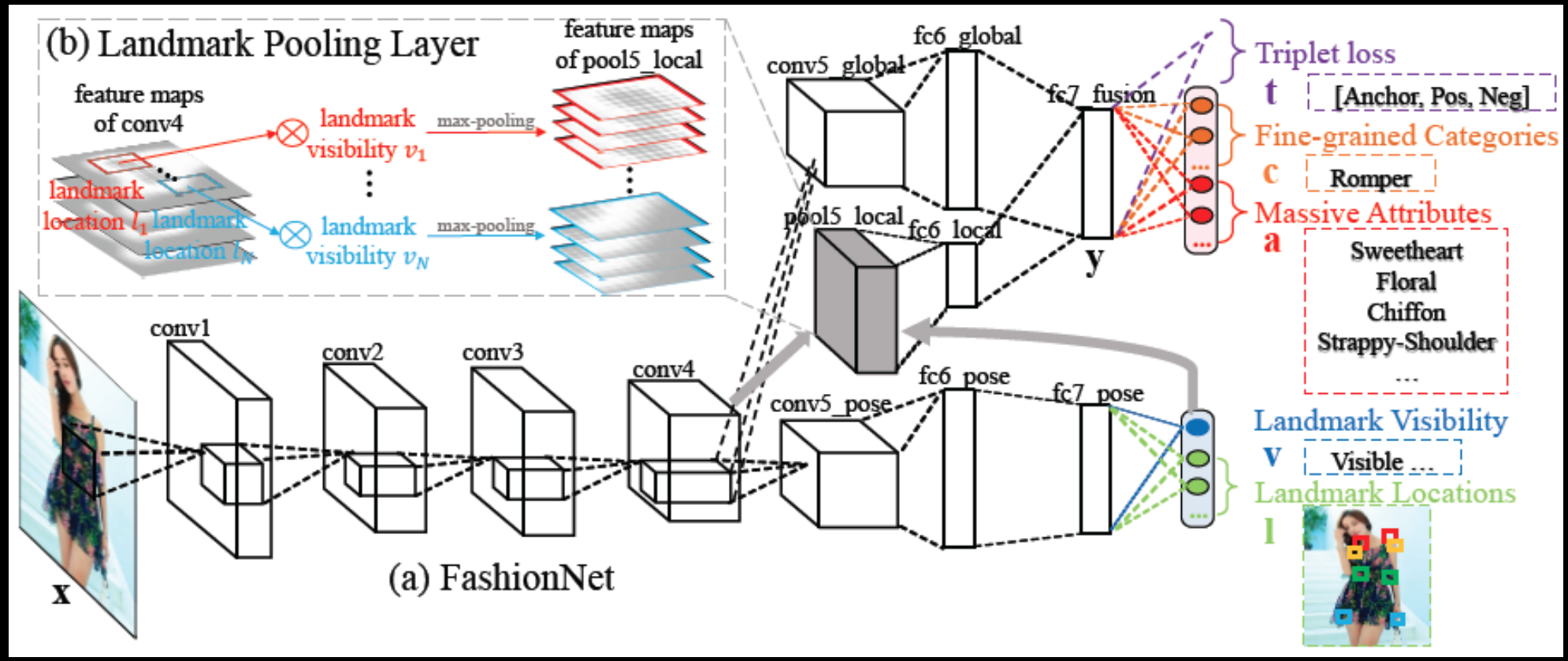
The interplay between identities and attributes

PID: 2000077658 (Forever21)

Ringer Tee (WOMEN)

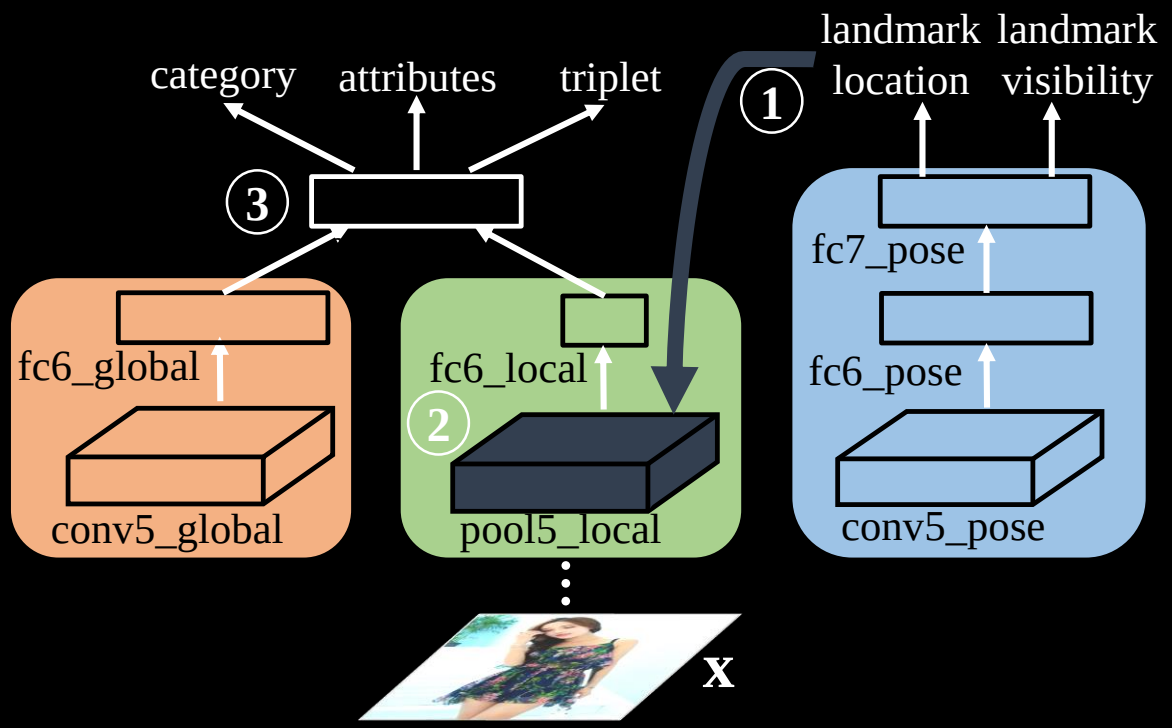
FashionNet

End-to-end System



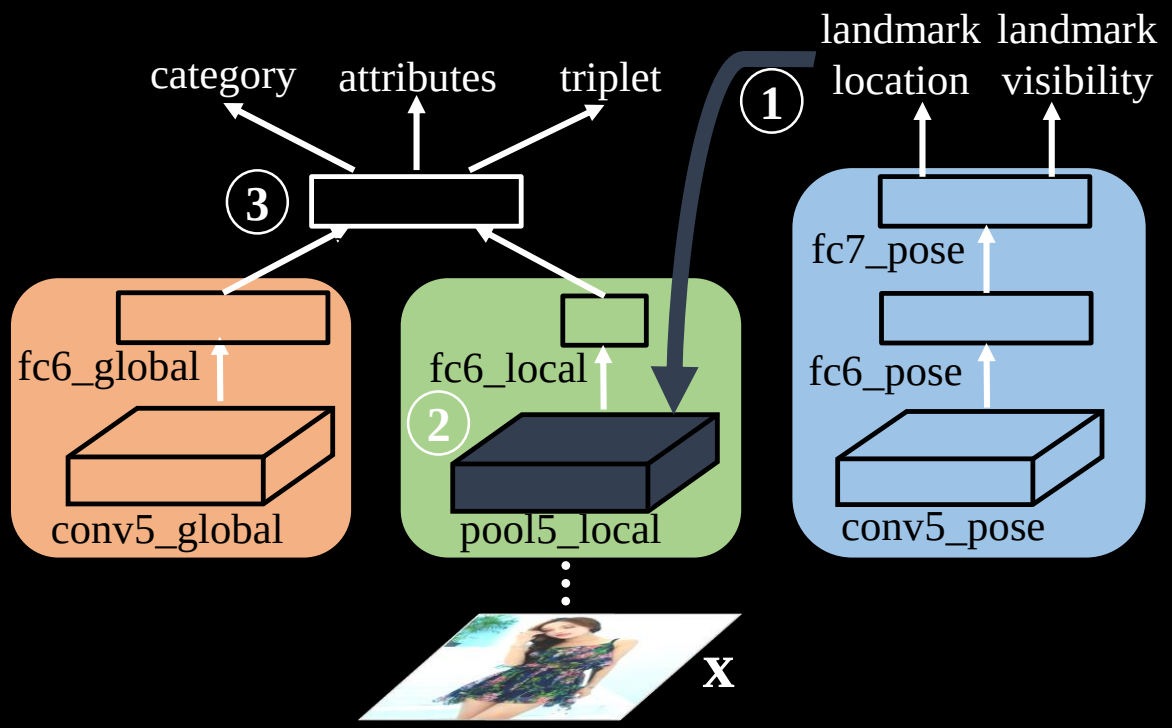
FashionNet

Forward Pass



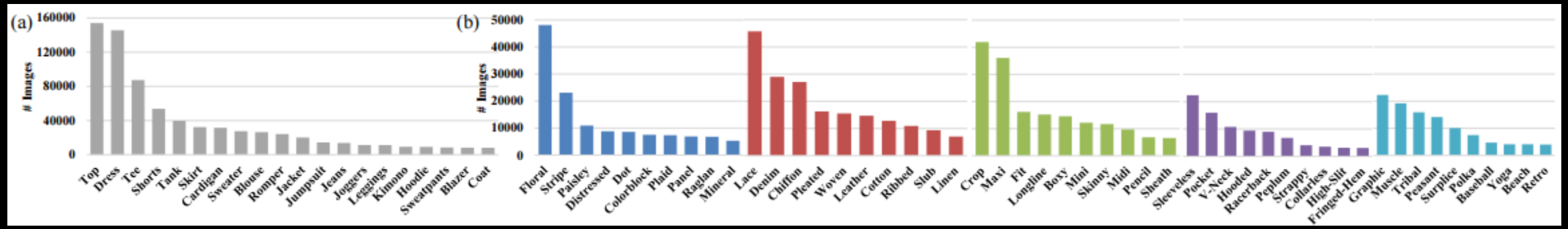
FashionNet

Backward Pass



Clothes Recognition

Attributes are noisy and imbalanced

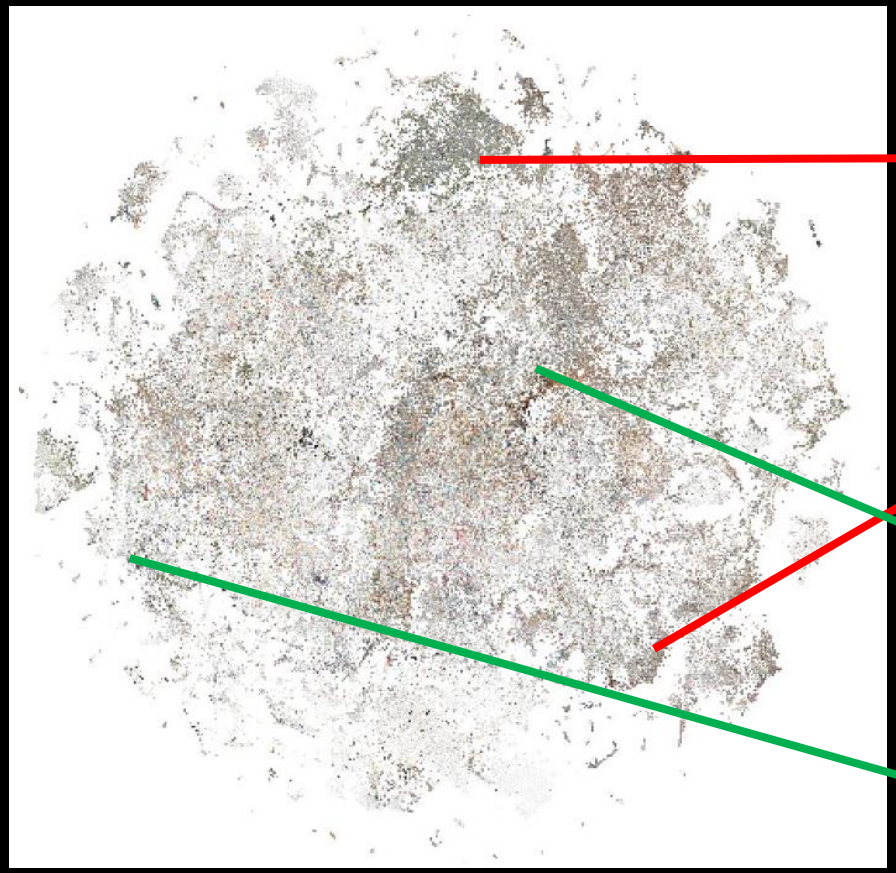


$$J = \sum_{i=1}^{n} \sum_{j=1}^{c_+} \sum_{k=1}^{c_-} \max(0, 1 - f_j(\boldsymbol{x}_i) + f_k(\boldsymbol{x}_i))$$

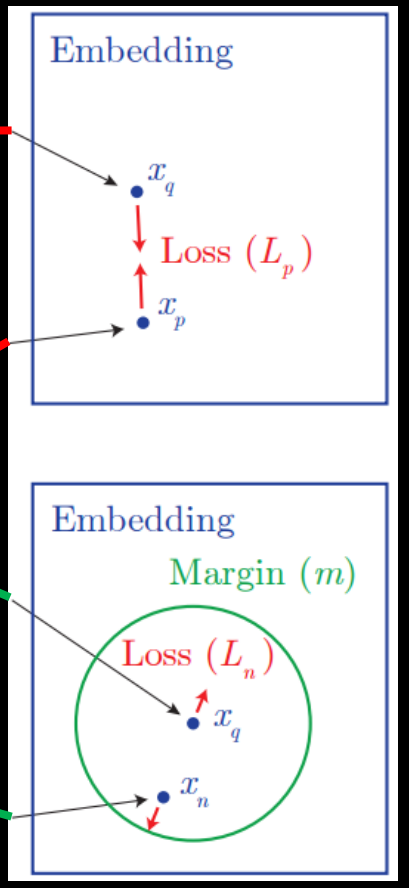
Multi-label Ranking Loss

Clothes Recognition

The number of identities are huge



Millions of fashion identities



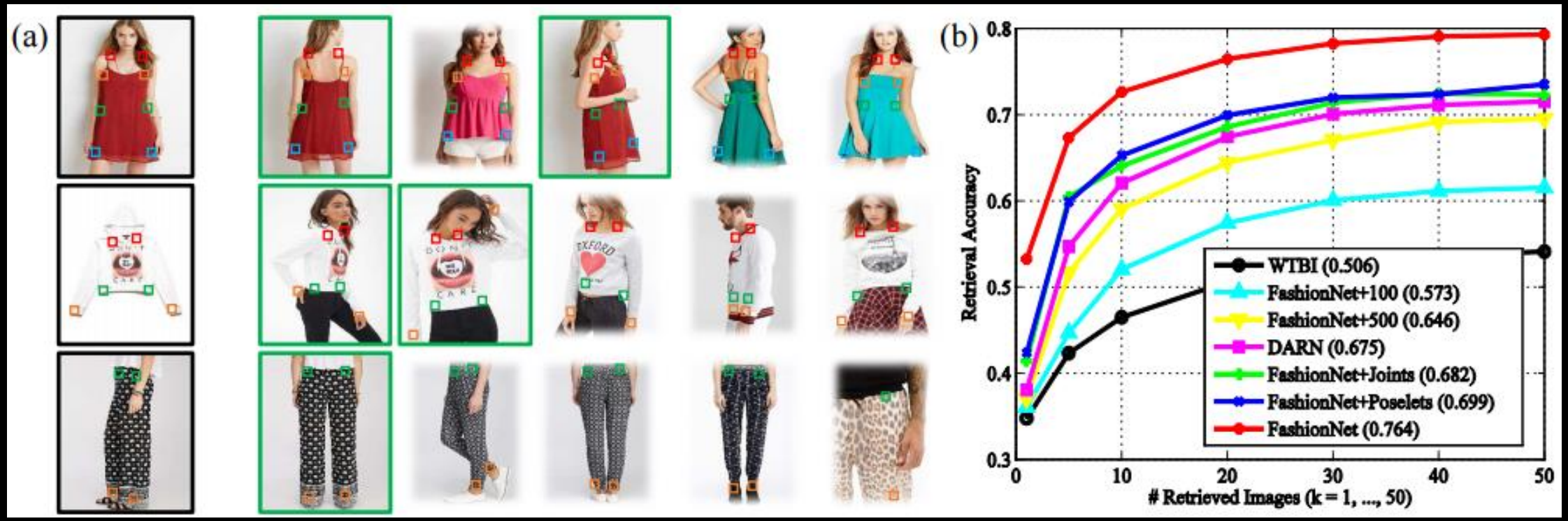
Positive Pair

Negative Pair

Hard Negative Mining

Clothes Recognition

In-shop Clothes Retrieval

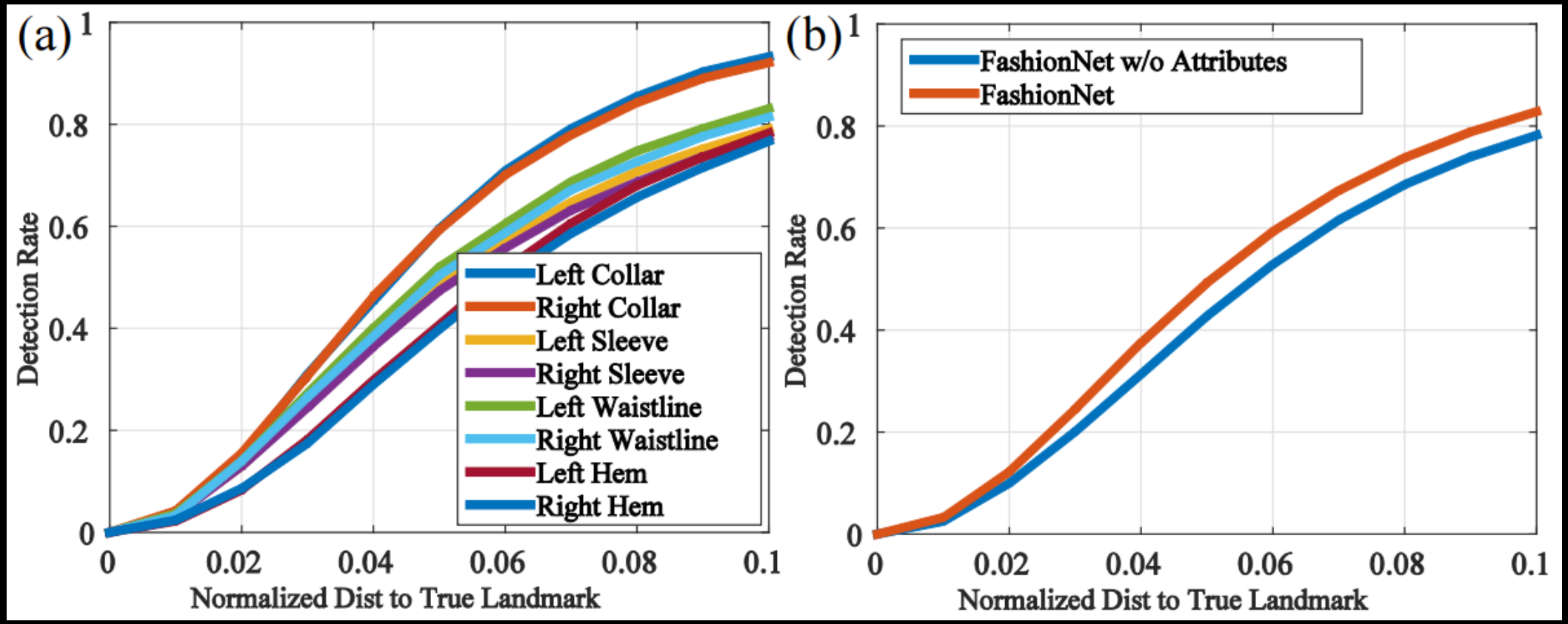


Clothes Recognition

Consumer-to-shop Clothes Retrieval

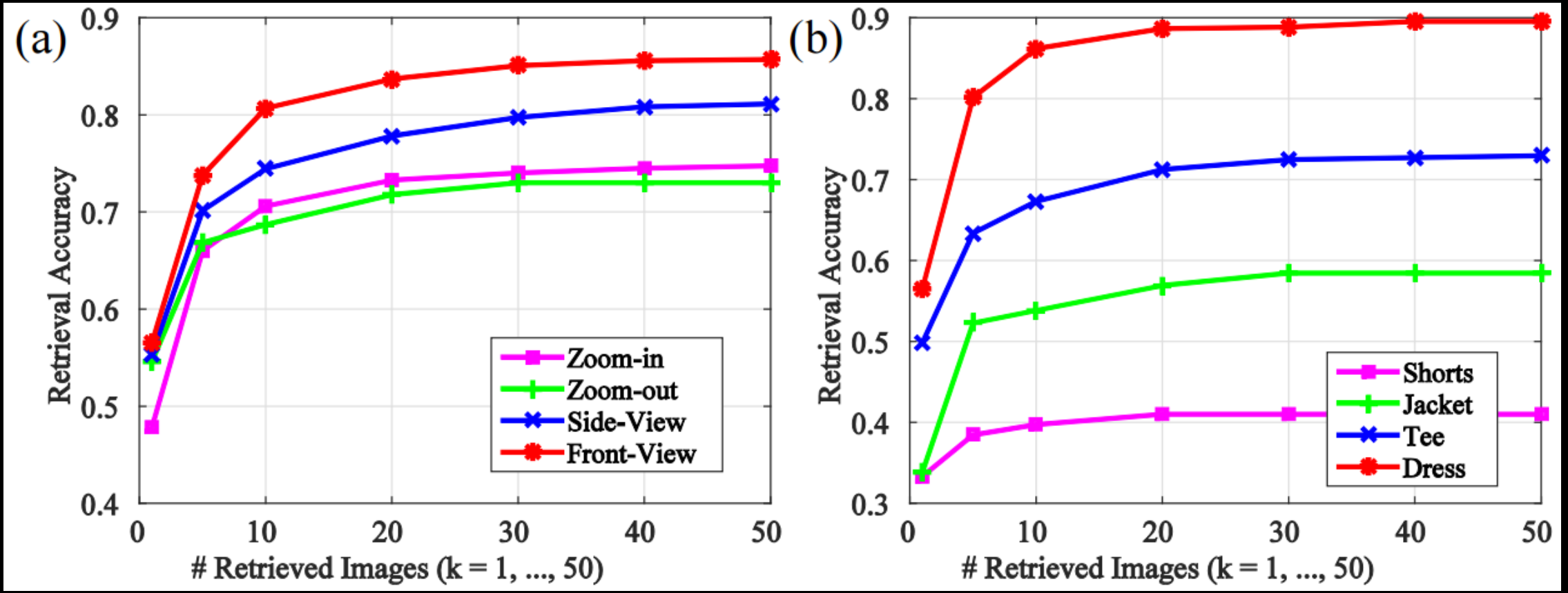
Clothes Recognition

Multi-task Learning



Clothes Recognition

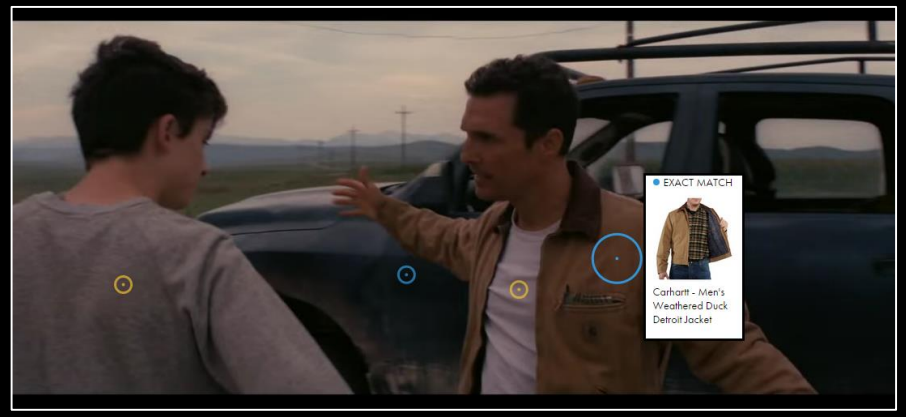
Further Analysis



Clothes Recognition

Applications

Similar Style Retrieval



Cloth Spotting in Video

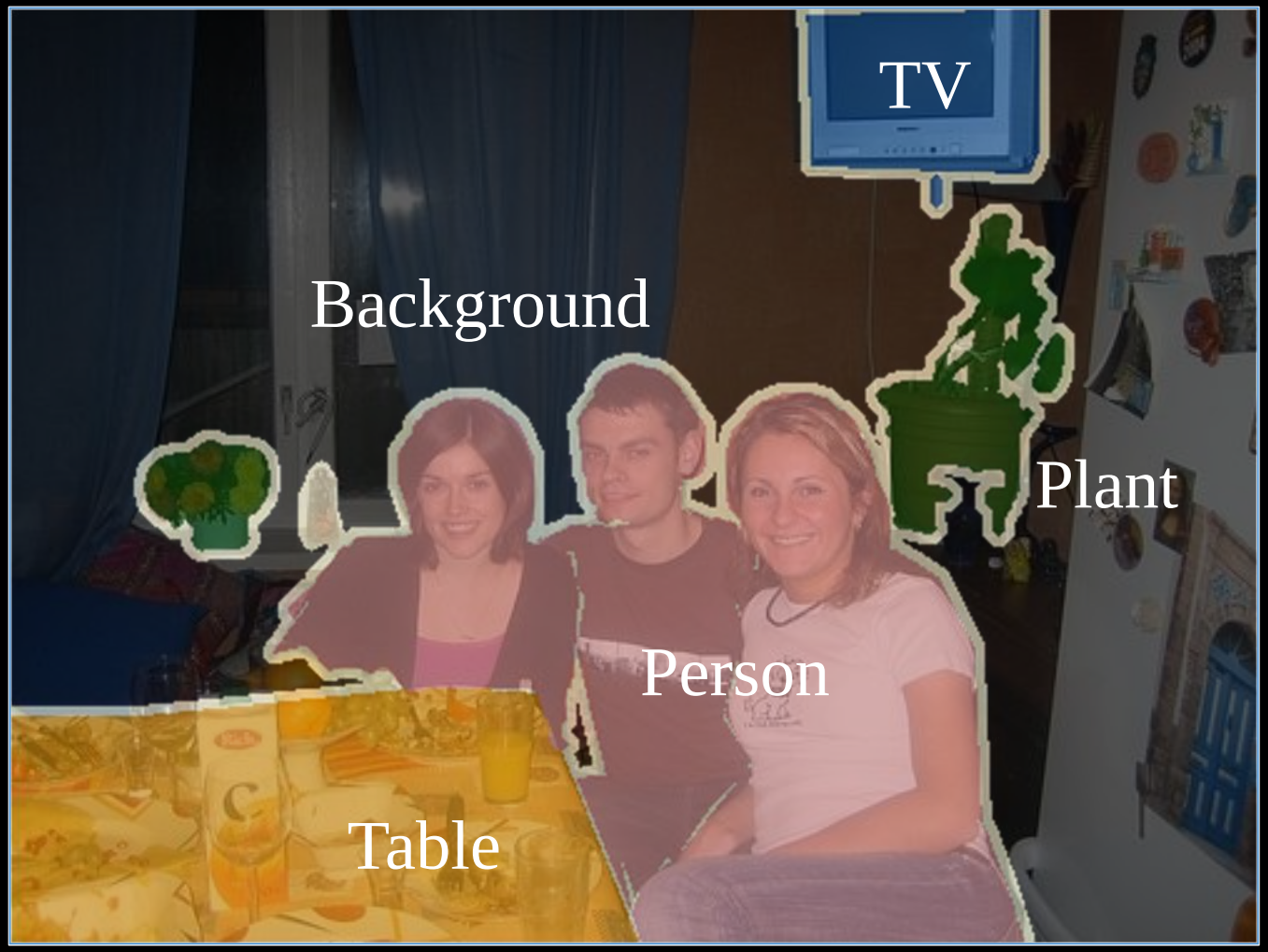
Street-to-shop

Fashion Assistant

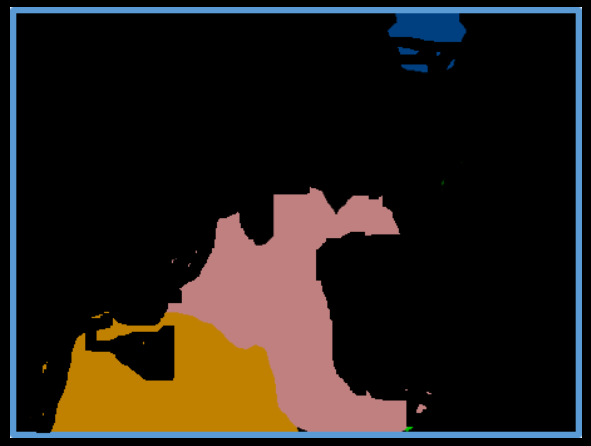
Part III: Deep Scene Understanding

Problem

Problem

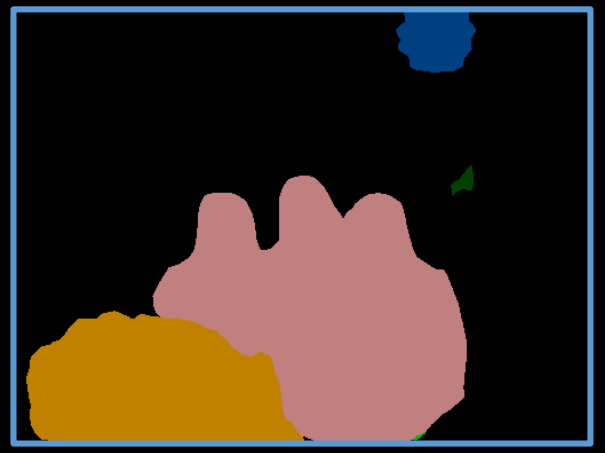


Previous Attempts

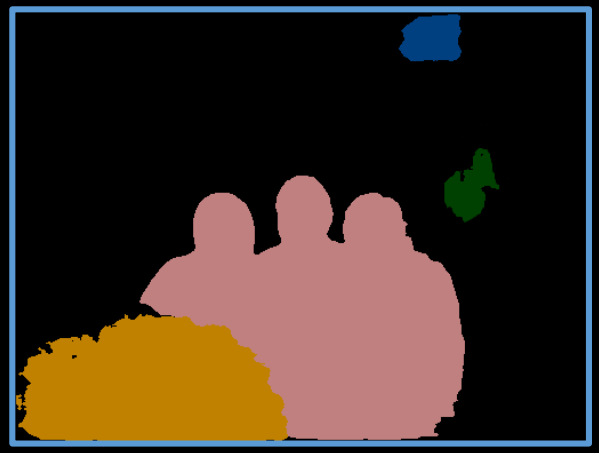


SVM

SVM + MRF

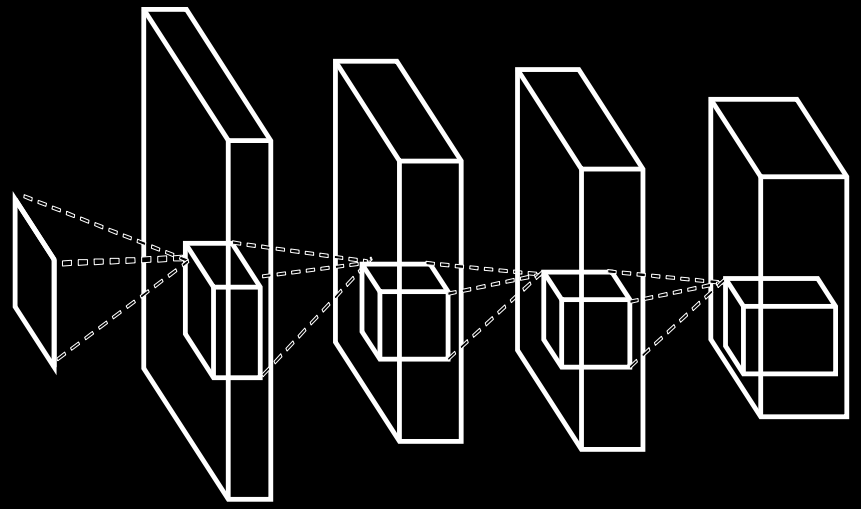


CNN



CNN + MRF ?

State-of-the-arts



Fully Convolutional Network

[Long et al. CVPR 2015]

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

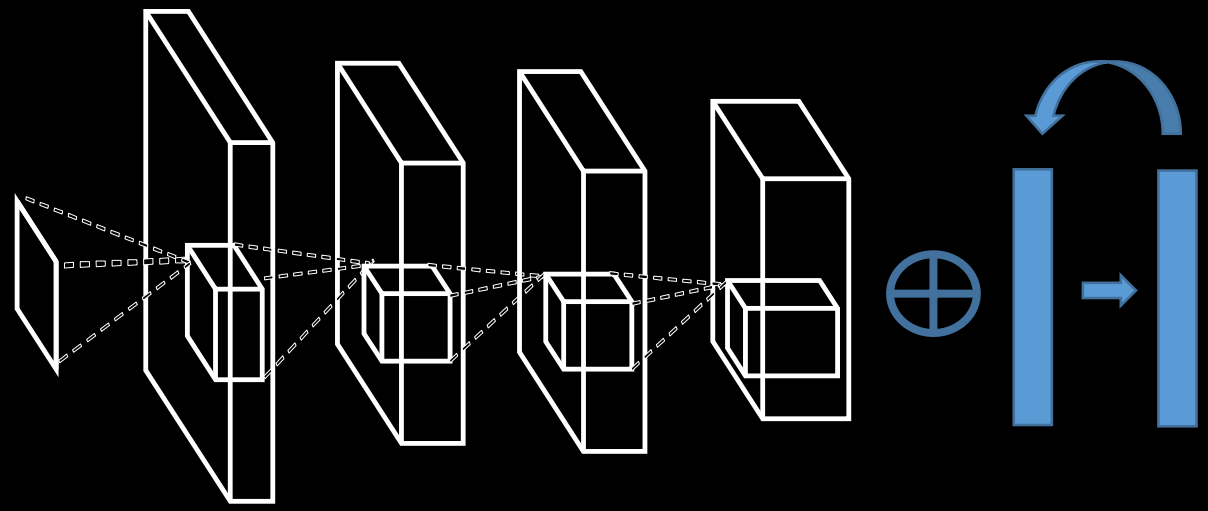
State-of-the-arts

DeepLab

[Chen et al. ICLR 2015]

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

State-of-the-arts

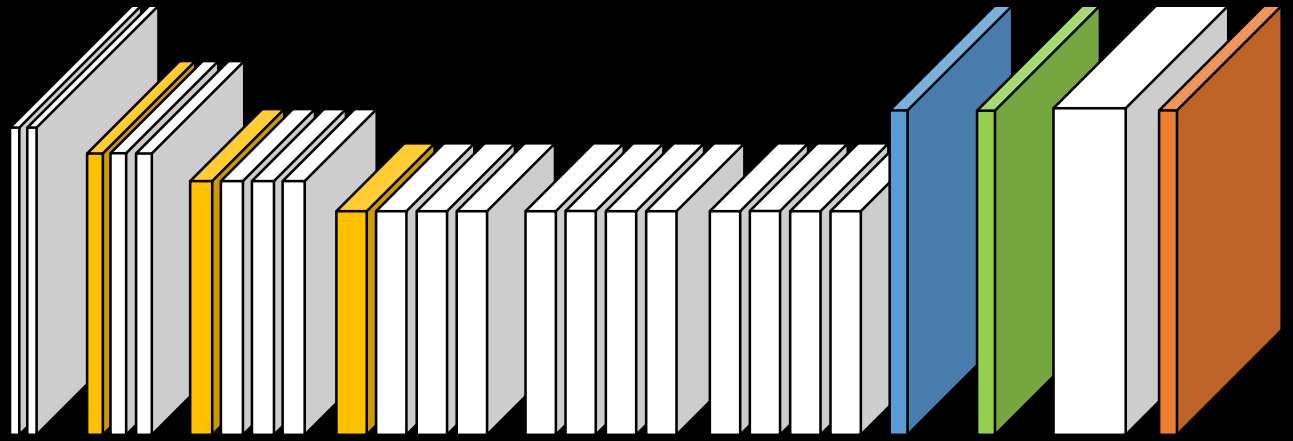


CRF as RNN

[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

Revisit MRF

$p_i(label =' table') = 0.8$

Energy Function

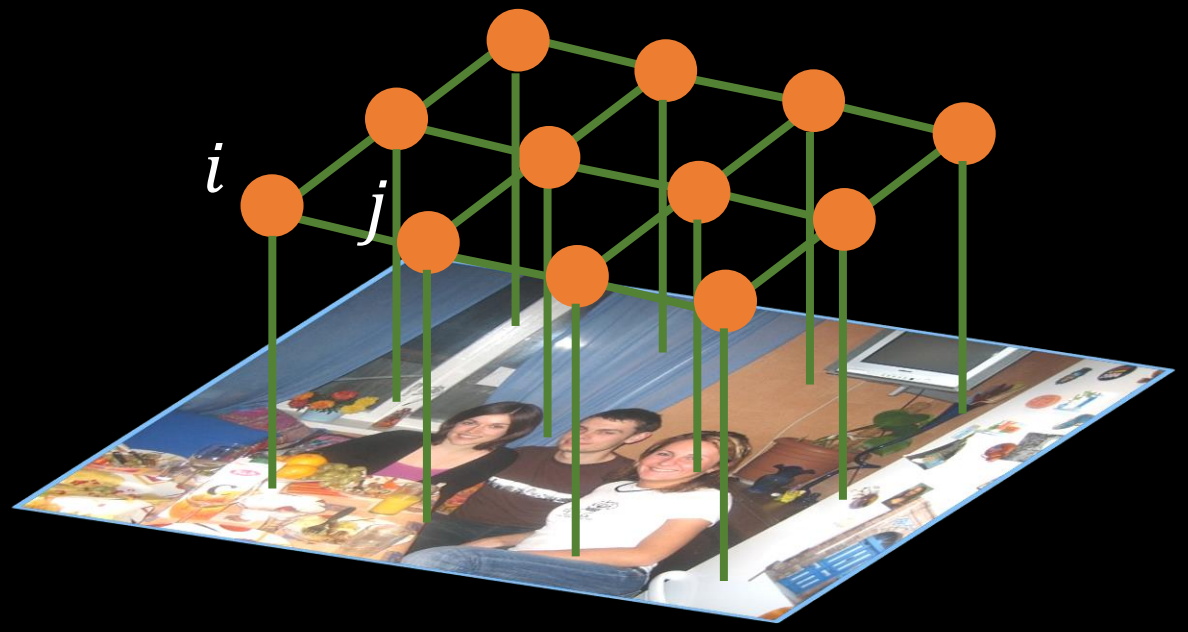
$$\min \quad E = Unary + Pair$$

Unary Term

$$Unary = -\sum_i \ln p_i(label)$$

Revisit MRF

$diss(i,j) = ($  ,  $) = 0.8$



Appearance Consistency

Energy Function

$$\min \quad E = Unary + Pair$$

Unary Term

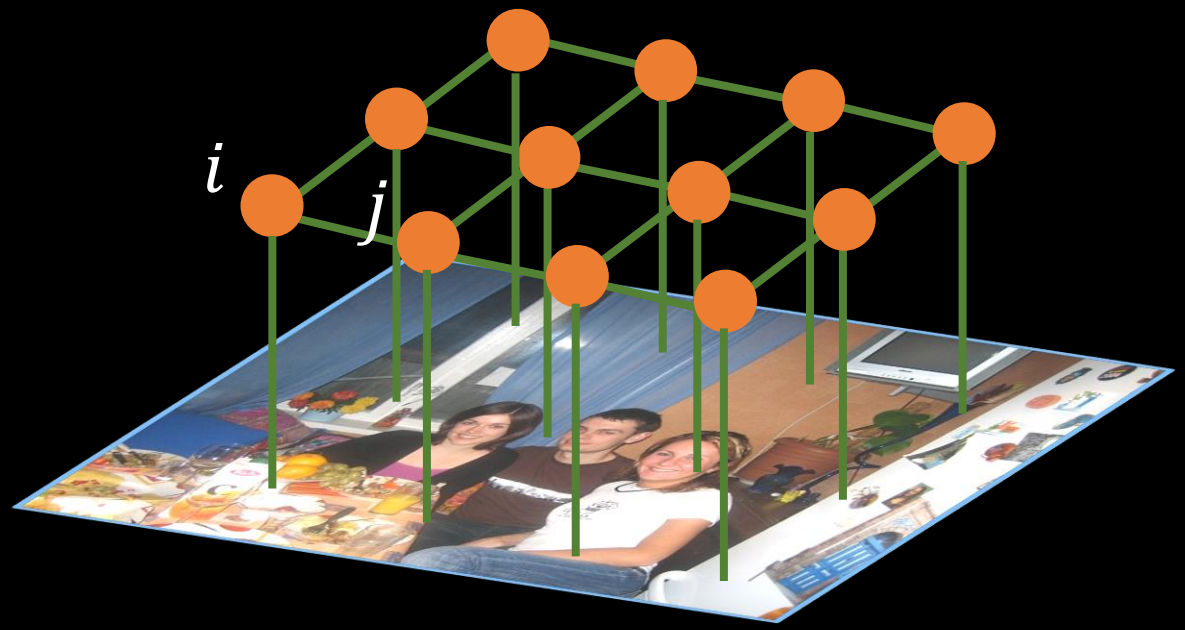
$$Unary = -\sum_i \ln p_i(label)$$

Pairwise Term

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

Revisit MRF

$cost(i; label =' table') = 0.1$



Label Consistency

Energy Function

$$\min \quad E = Unary + Pair$$

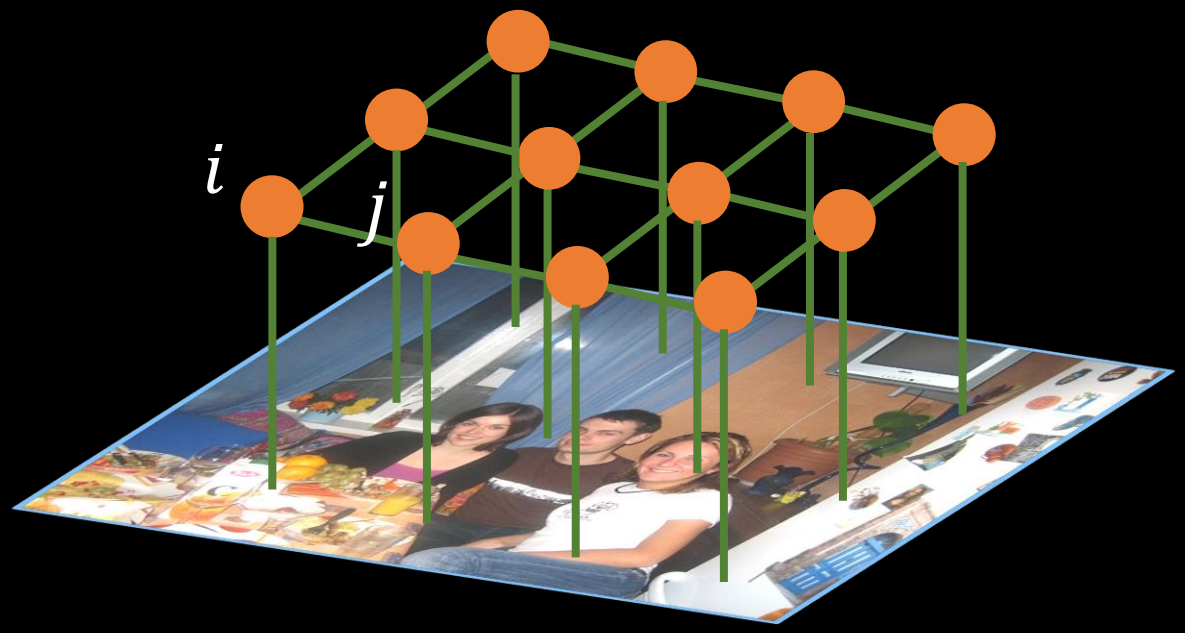
Unary Term

$$Unary = -\sum_{i} \ln p_i(label)$$

Pairwise Term

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

Richer Relationships in DPN



Energy Function

$$\min \quad E = Unary + Pair$$

Unary Term

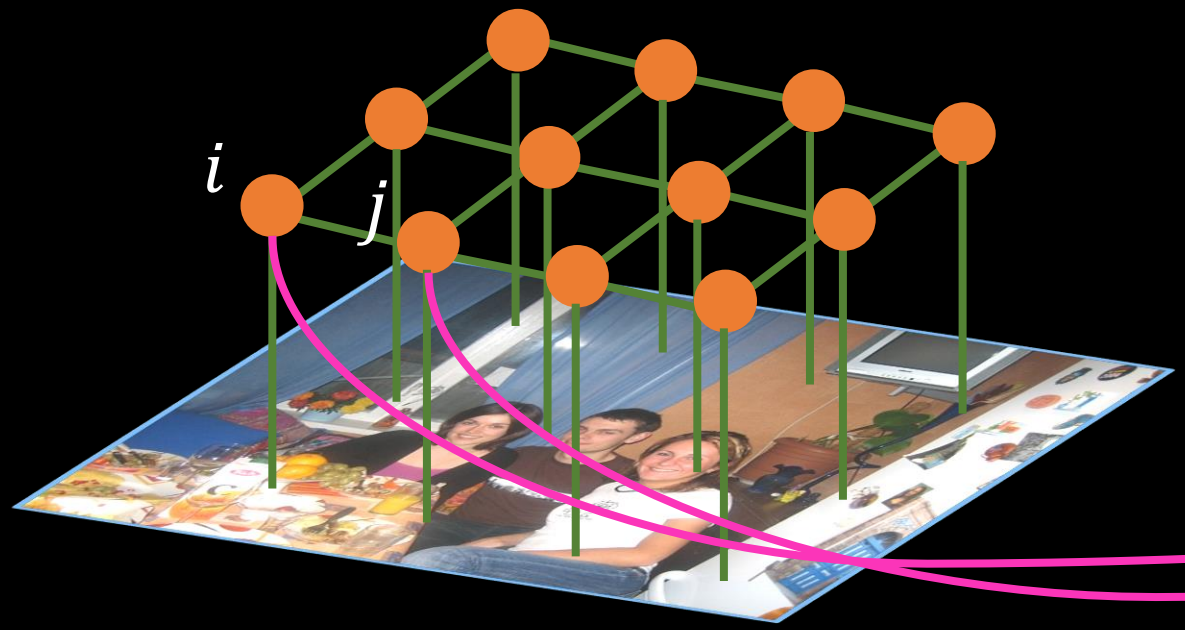
$$Unary = -\sum_{i} \ln p_i(label)$$

Pairwise Term

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

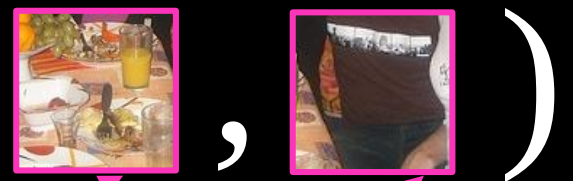
Richer Relationships in DPN

Triple Penalty



Pairwise Term

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

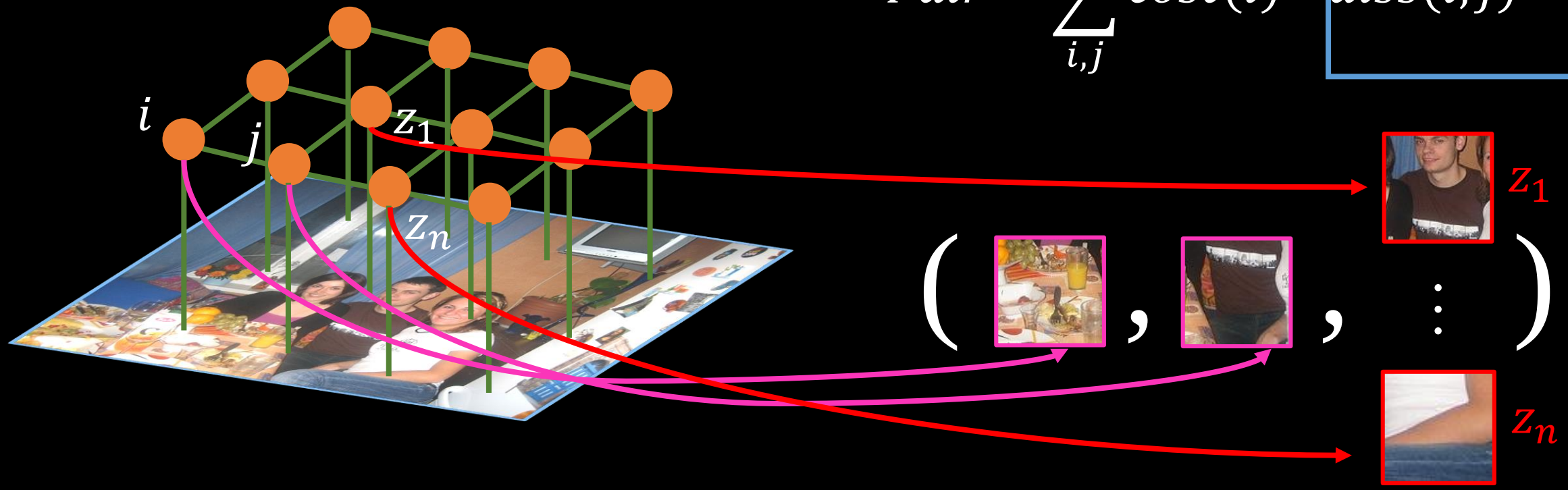


Richer Relationships in DPN

Triple Penalty

Pairwise Term

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

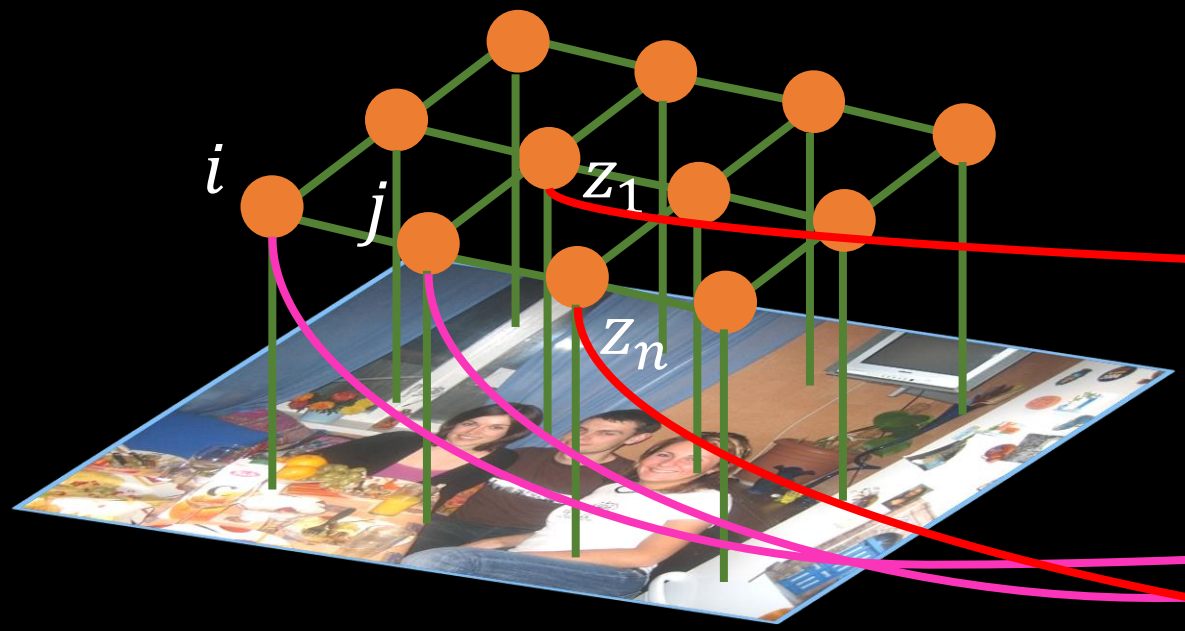


Richer Relationships in DPN

Triple Penalty

Pairwise Term

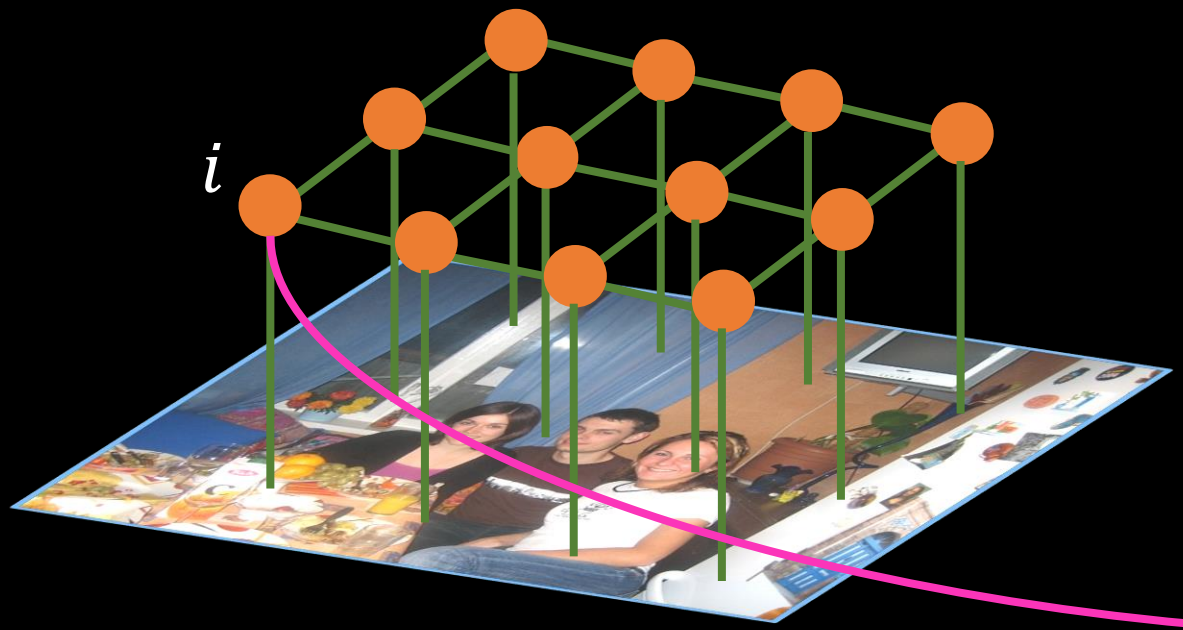
$$Pair = \sum_{i,j} cost(i) * \sum_{z} diss(i, j; z)$$



Triple Penalty

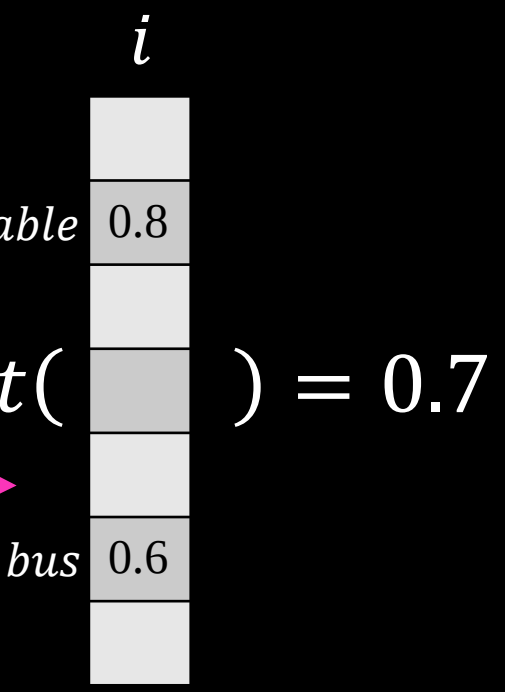
Richer Relationships in DPN

Mixture of Label Contexts



Pairwise Term

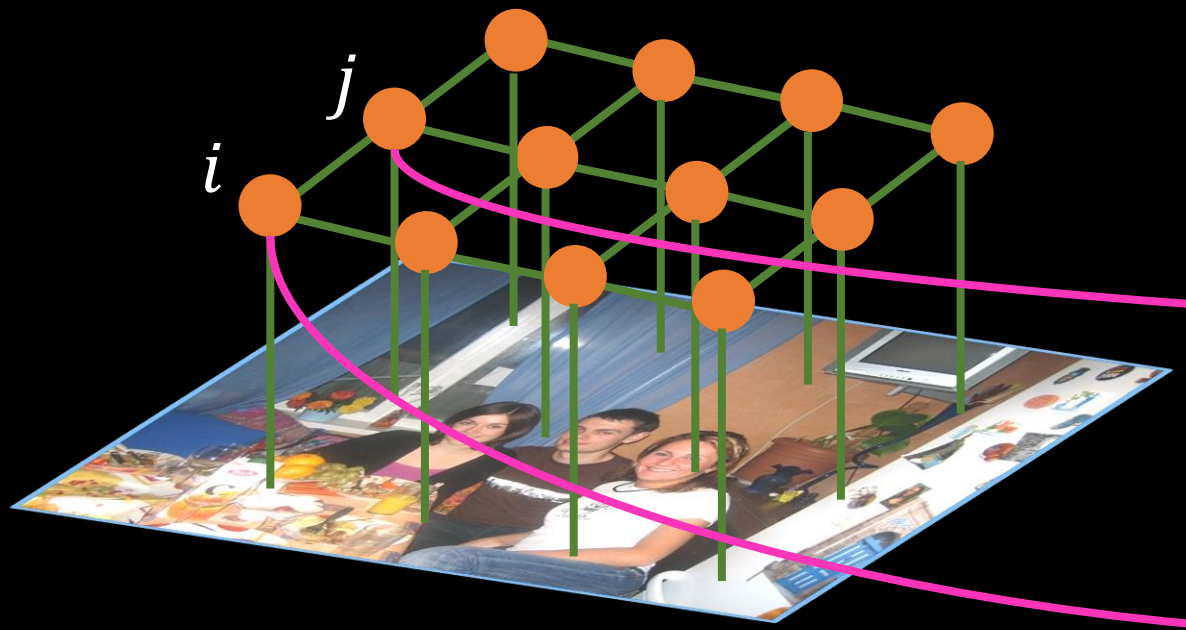
$$Pair = \sum_{i,j} cost(i) * \sum_{z} diss(i,j;z)$$



$$cost(\quad) = 0.7$$

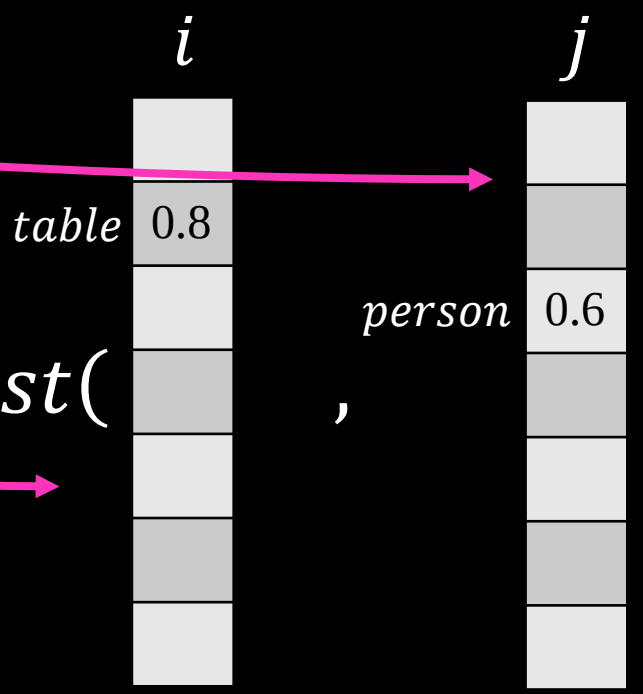
Richer Relationships in DPN

Mixture of Label Contexts



Pairwise Term

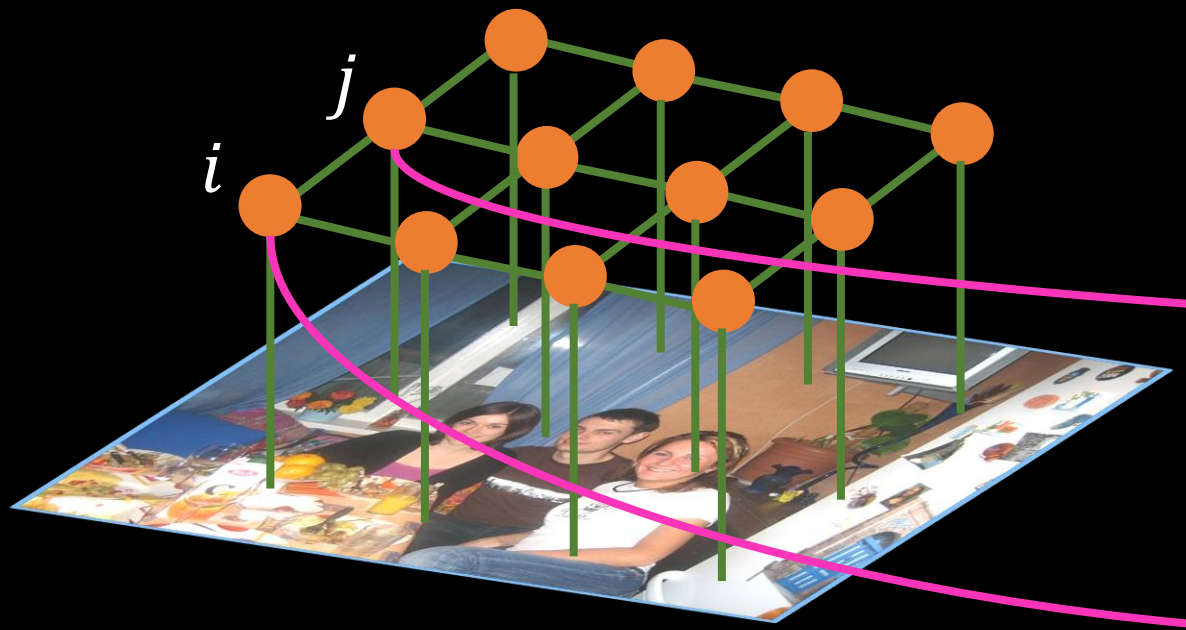
$$Pair = \sum_{i,j} cost(i,j) * \sum_{z} diss(i,j;z)$$



$$cost(\quad , \quad) = 0.2$$

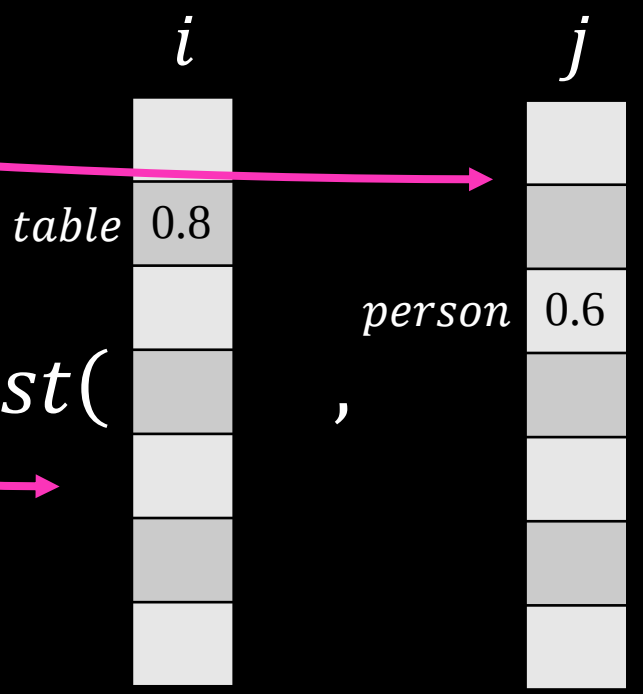
Richer Relationships in DPN

Mixture of Label Contexts



Pairwise Term

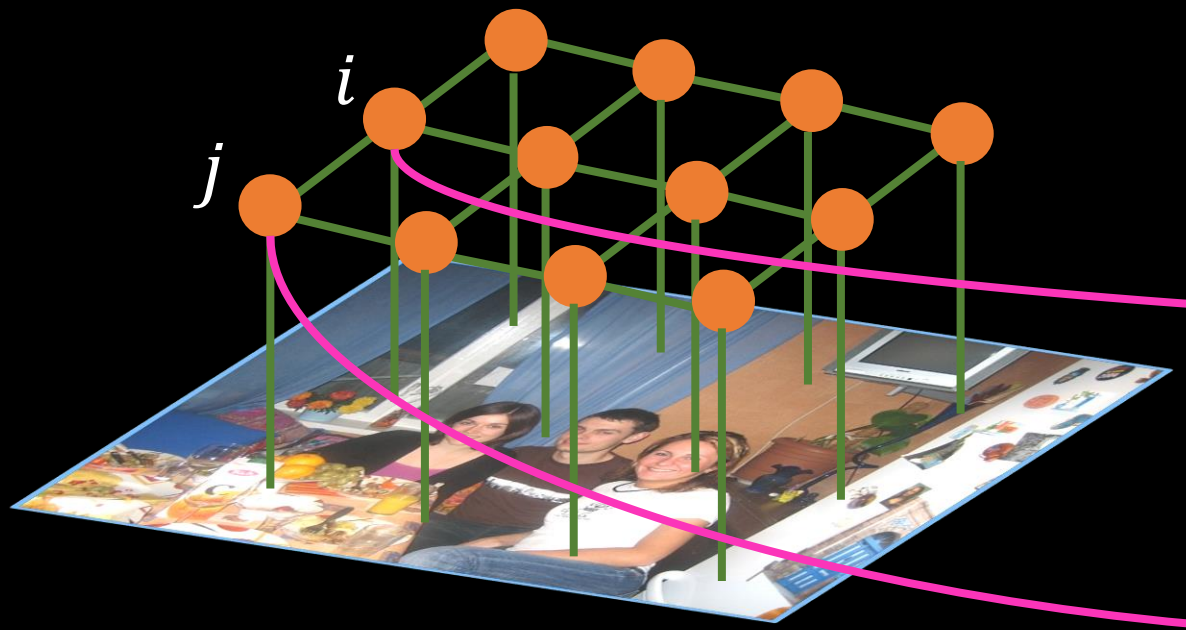
$$Pair = \sum_{i,j} cost(i,j) * \sum_{z} diss(i,j;z)$$



$$cost(\quad , \quad) = 0.2$$

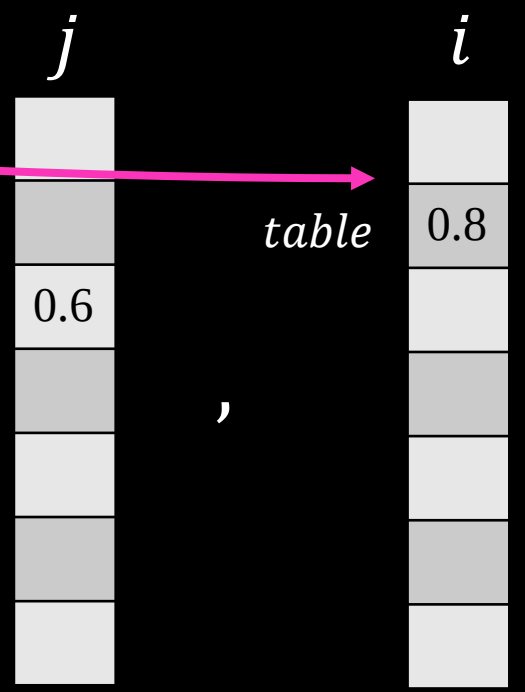
Richer Relationships in DPN

Mixture of Label Contexts



Pairwise Term

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

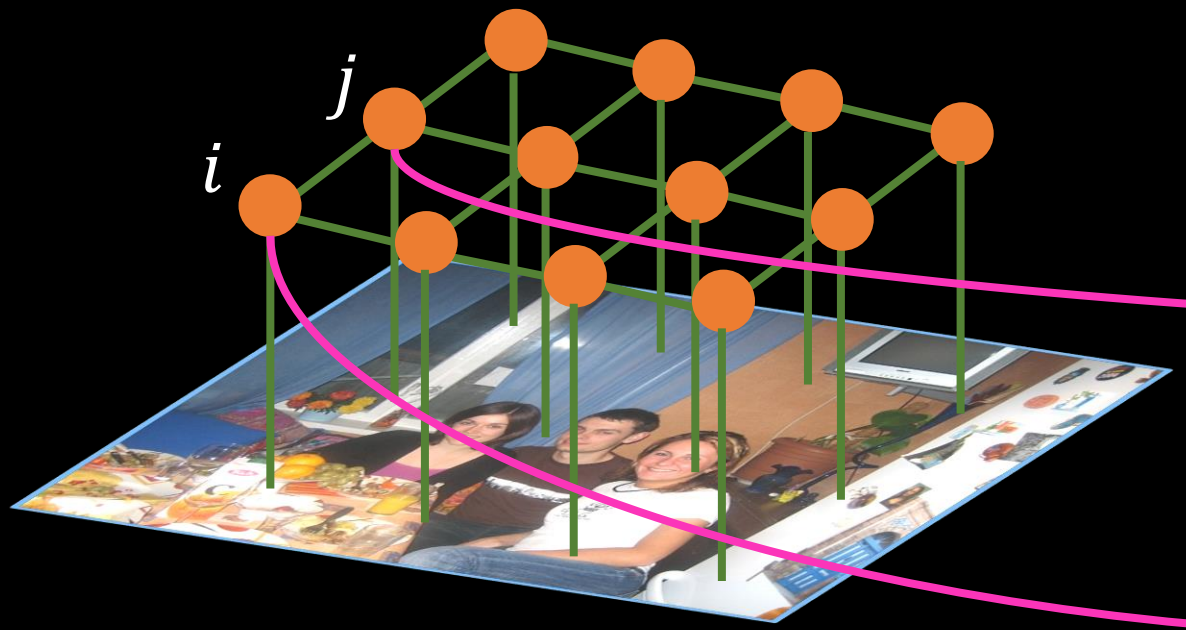


$$cost(\ ,\ ) = 0.8$$

Spatial Order

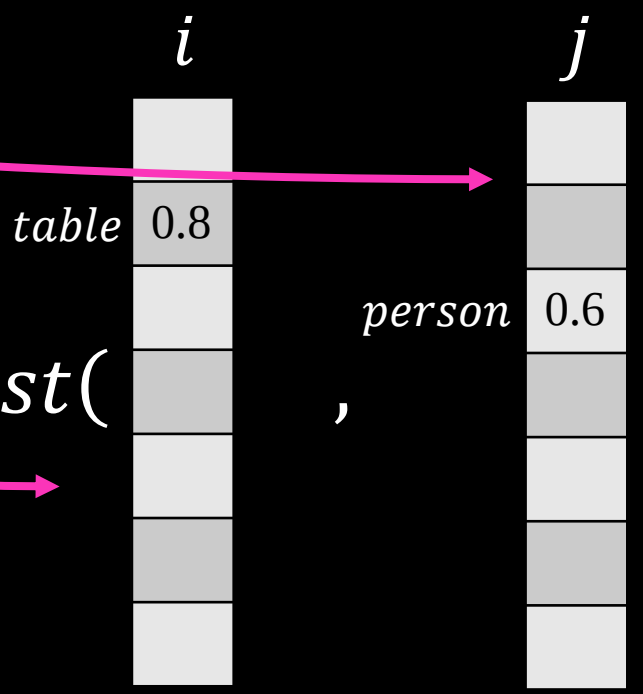
Richer Relationships in DPN

Mixture of Label Contexts



Pairwise Term

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



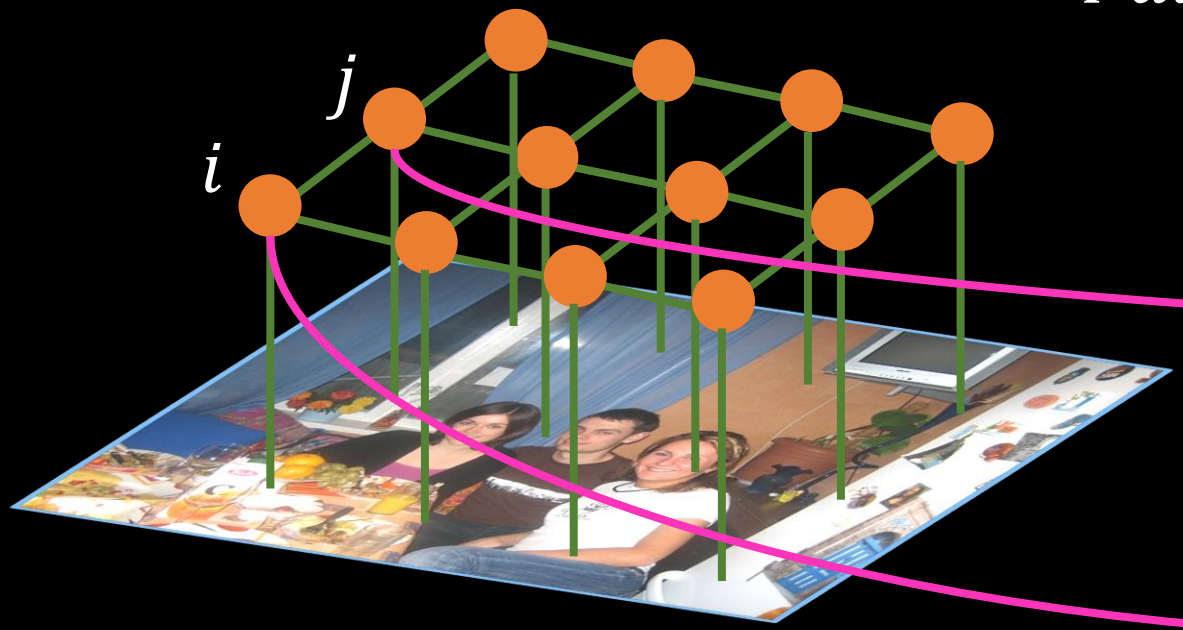
$$cost(\quad , \quad) = 0.2$$

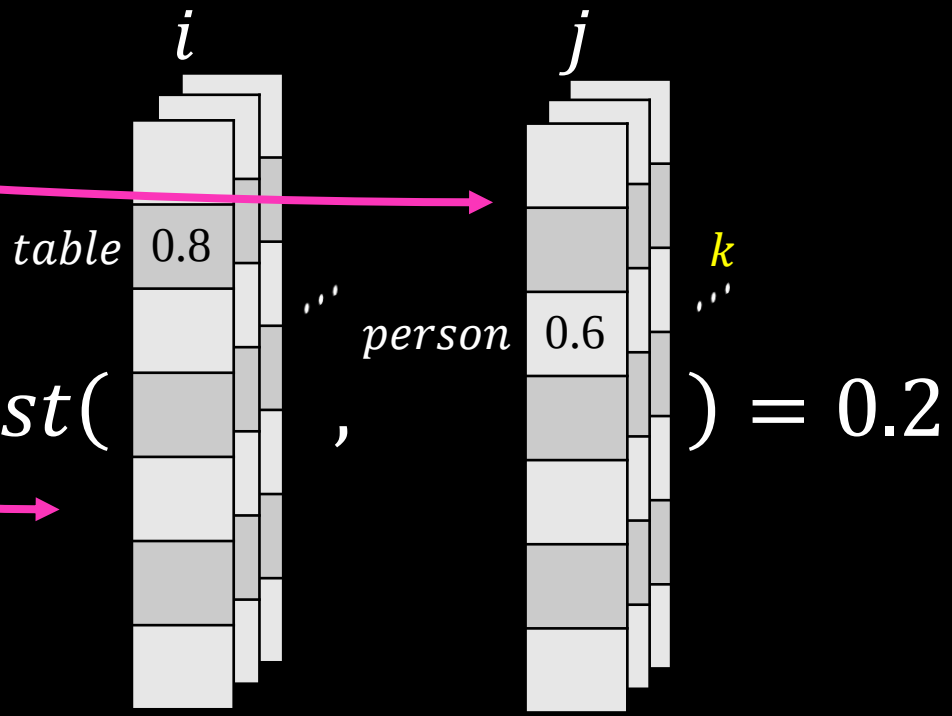
Richer Relationships in DPN

Mixture of Label Contexts

Pairwise Term

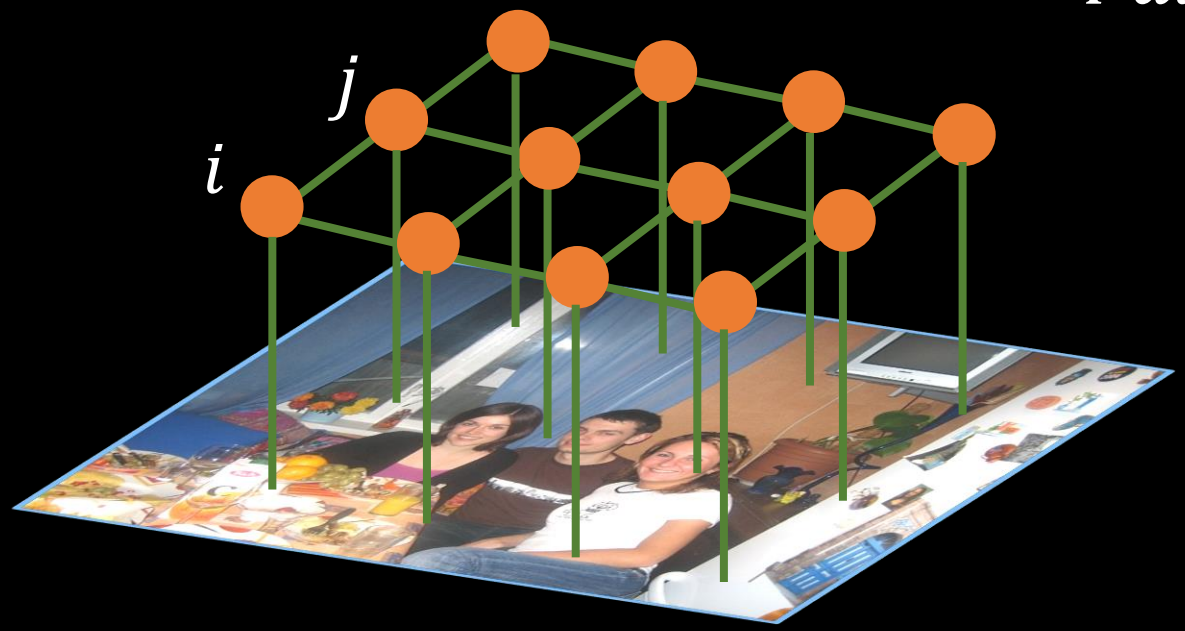
$$Pair = \sum_{i,j} \boxed{\sum_{k} cost_k(i,j)} * \sum_{z} diss(i,j;z)$$





Mixture of Label Contexts

Solve High-order MRF as Convolution



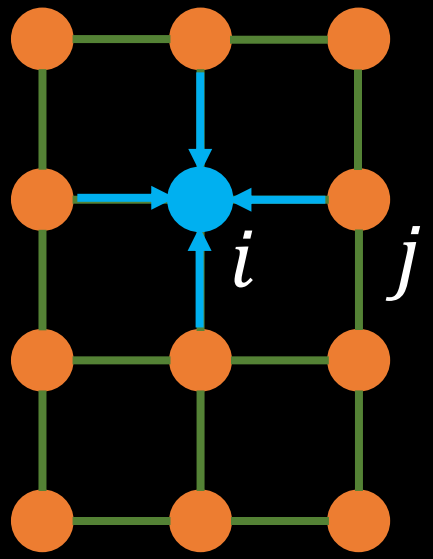
Pairwise Term

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

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



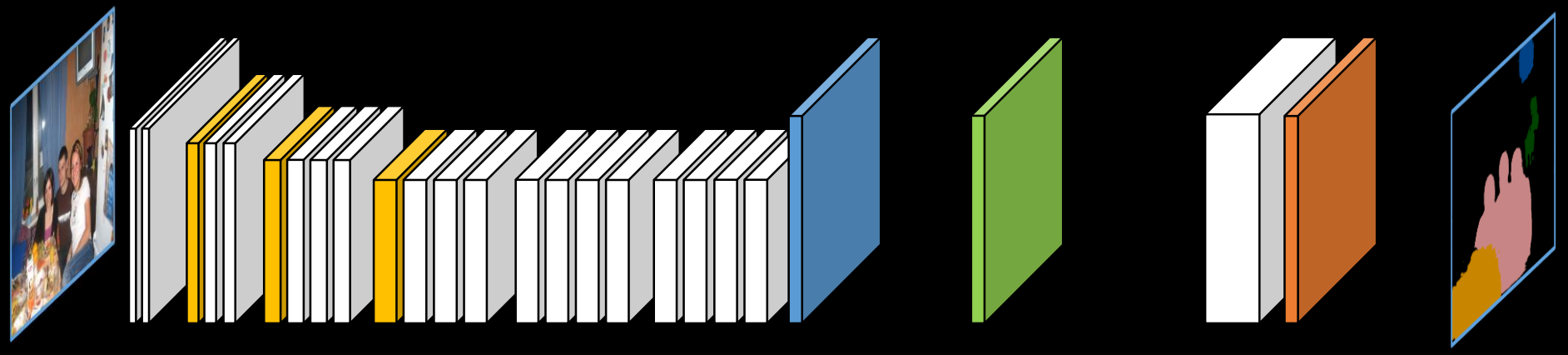
Iterative Updating Formula

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

Summation Convolution

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

Deep Parsing Network



Unary Term

Pairwise Term

Convolution

Deconvolution

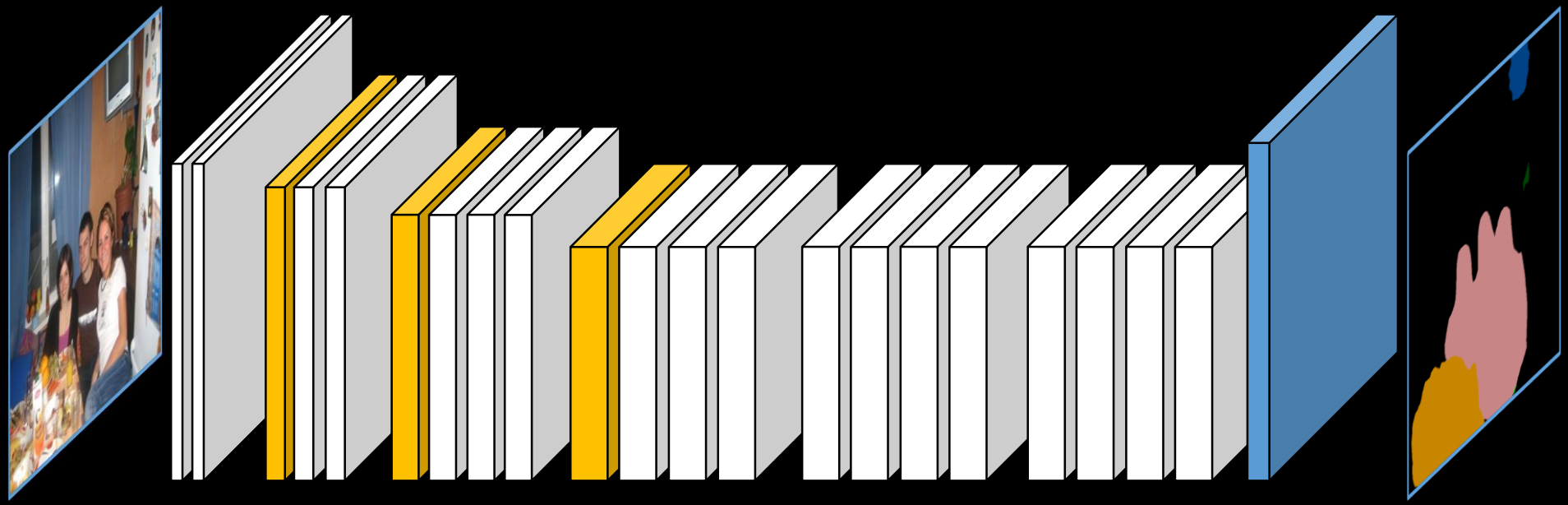
Local Convolution

Max Pooling

Min Pooling

Deep Parsing Network

Unary Term



Fine-tuned VGG-16 Network

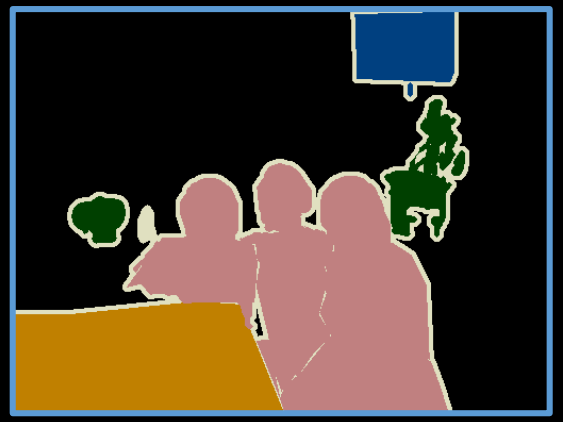
Convolution

Max Pooling

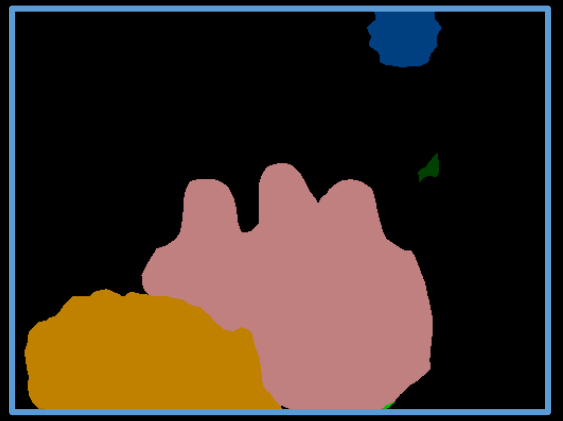
Deconvolution

Deep Parsing Network

Original Image

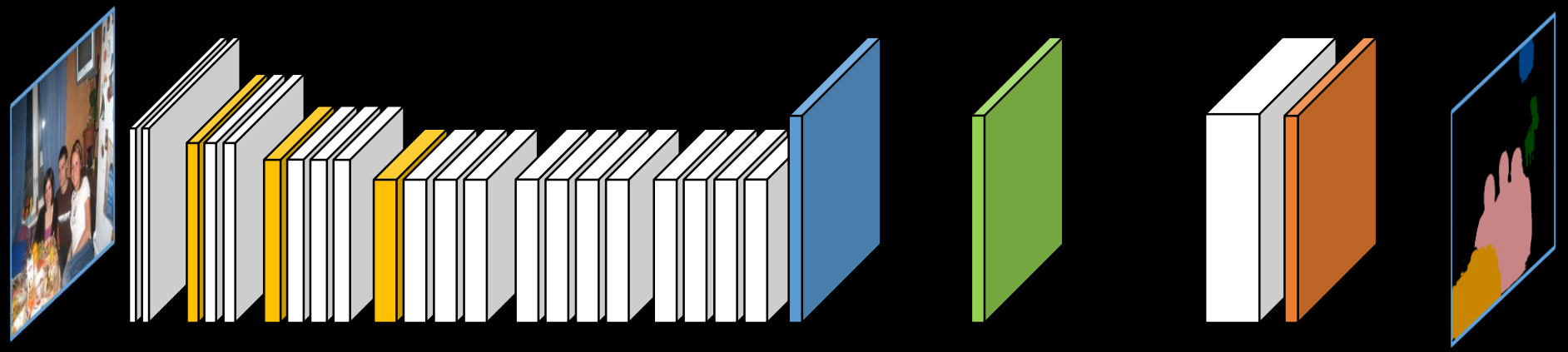


Ground Truth



Unary Term

Deep Parsing Network



Convolution

Deconvolution

Local Convolution

Max Pooling

Min Pooling

Deep Parsing Network

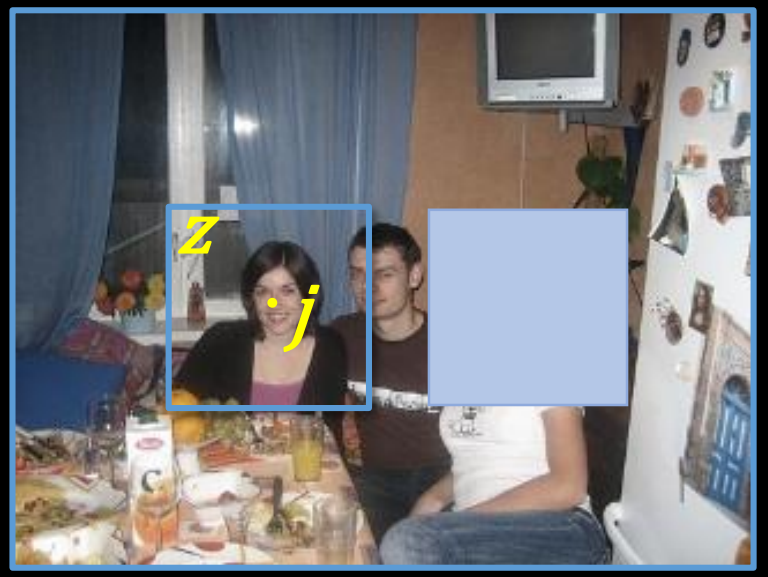
Triple Penalty

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

Deep Parsing Network

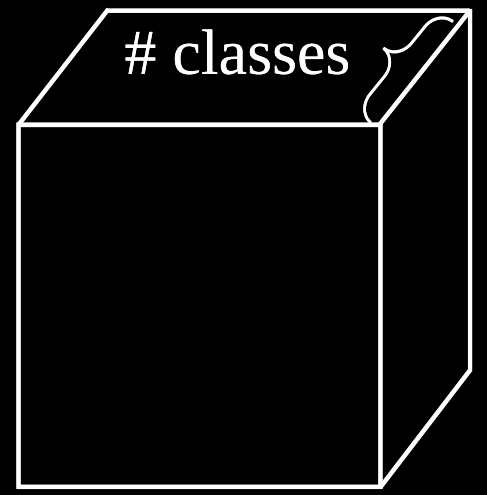
Triple Penalty

$$\sum_{z} diss(j; z) * p_z$$



j

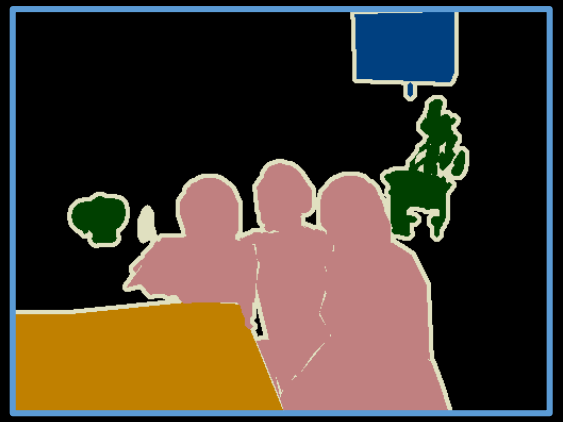
Local Conv



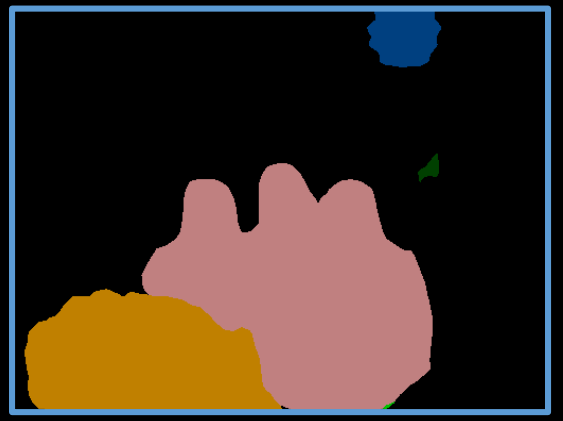
Unary Term

Deep Parsing Network

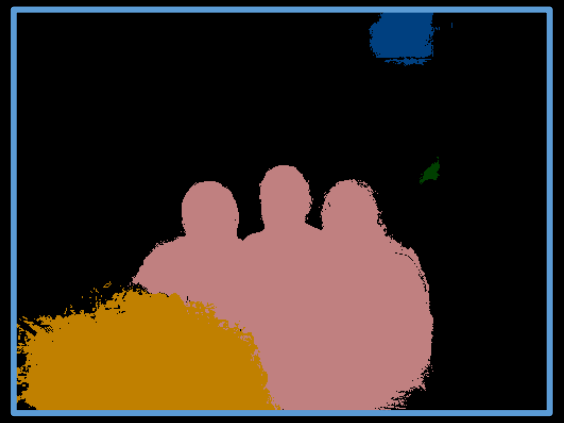
Original Image



Ground Truth

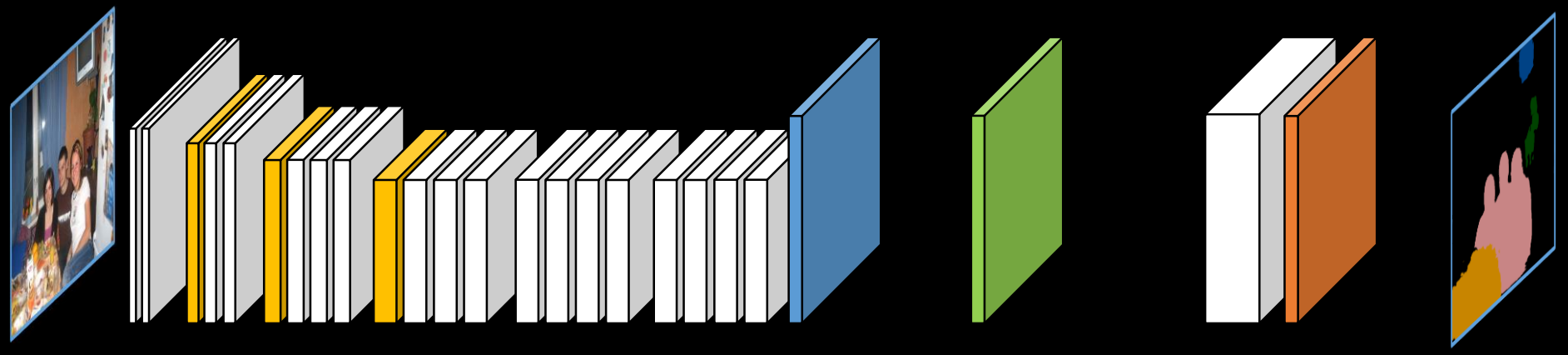


Unary Term



Triple Penalty

Deep Parsing Network



Unary Term

Pairwise Term

Convolution

Deconvolution

Local Convolution

Max Pooling

Min Pooling

Deep Parsing Network

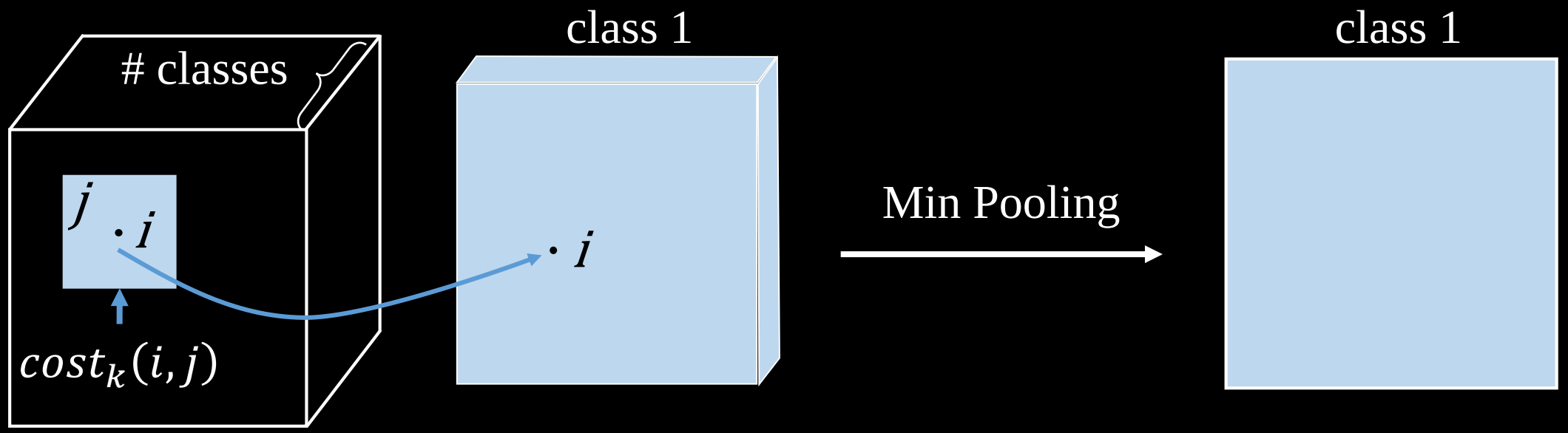
Mixture of Label Contexts

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

Deep Parsing Network

Mixture of Label Contexts

Triple Penalty Result

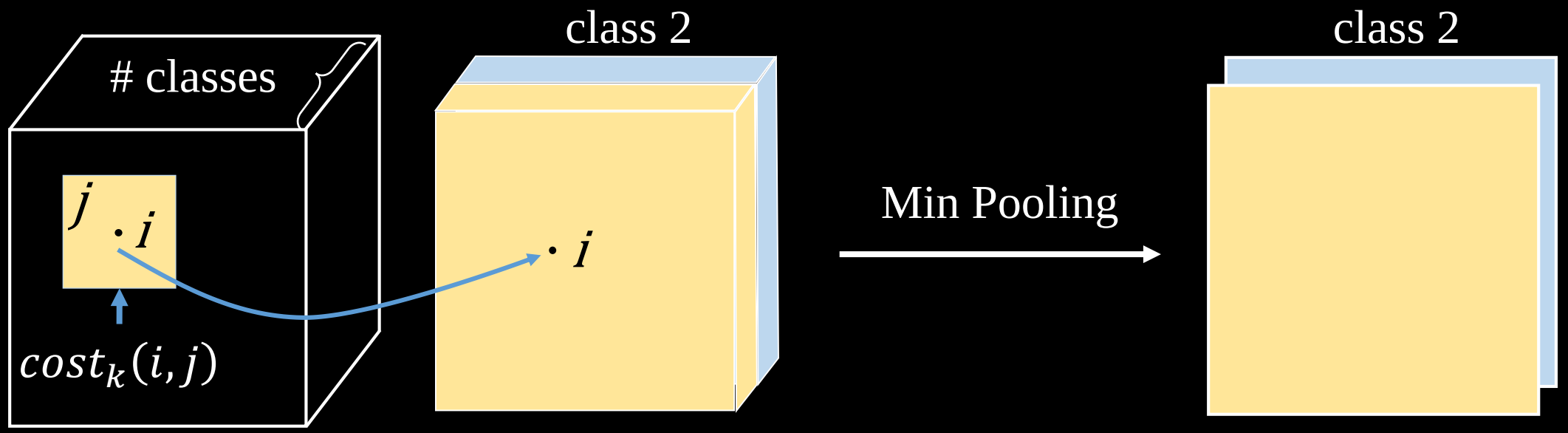


tri $\sum_j cost_k(i,j) * tri(j)$ $\sum_j \sum_k cost_k(i,j) * tri(j)$

Deep Parsing Network

Mixture of Label Contexts

Triple Penalty Result

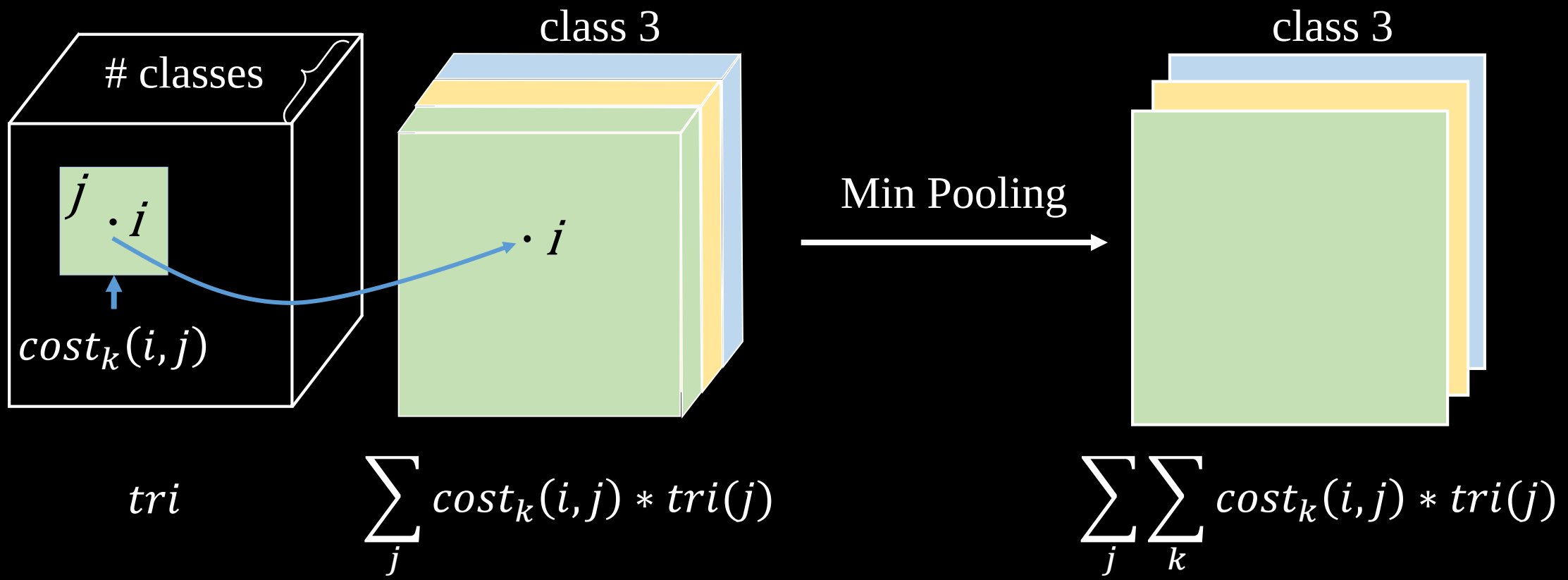


tri $\sum_j cost_k(i,j) * tri(j)$ $\sum_j \sum_k cost_k(i,j) * tri(j)$

Deep Parsing Network

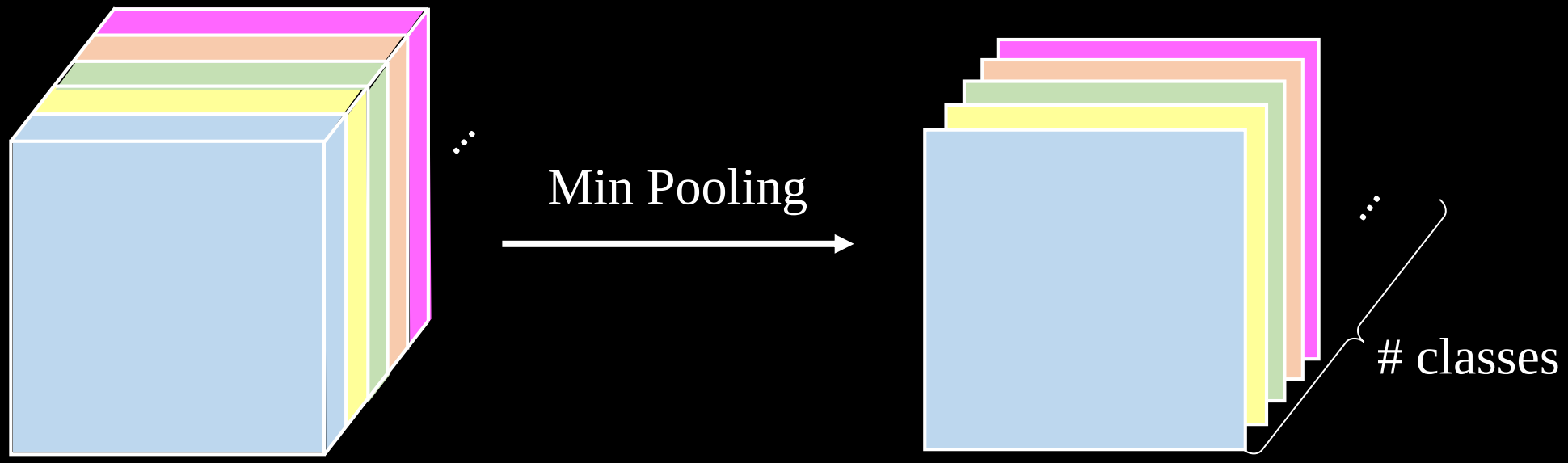
Mixture of Label Contexts

Triple Penalty Result



Deep Parsing Network

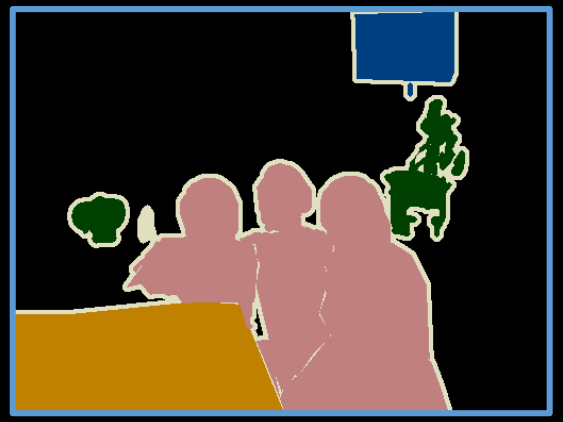
Mixture of Label Contexts



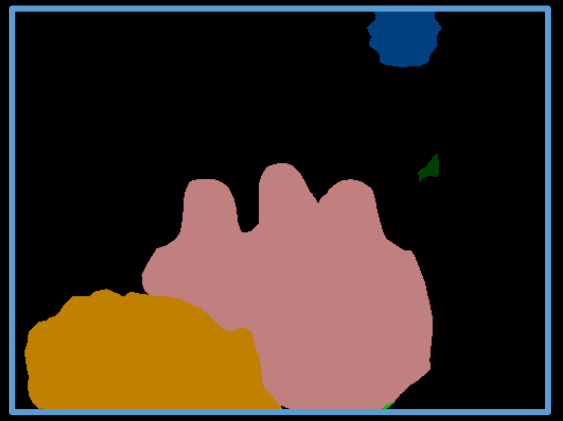
$$\sum_{j}\sum_{k} cost_k(i,j) * tri(j)$$

Deep Parsing Network

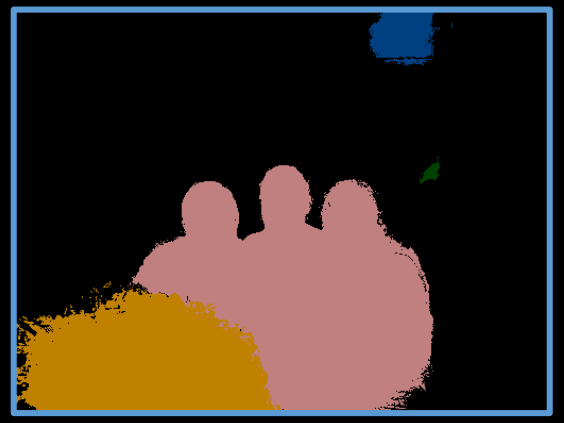
Original Image



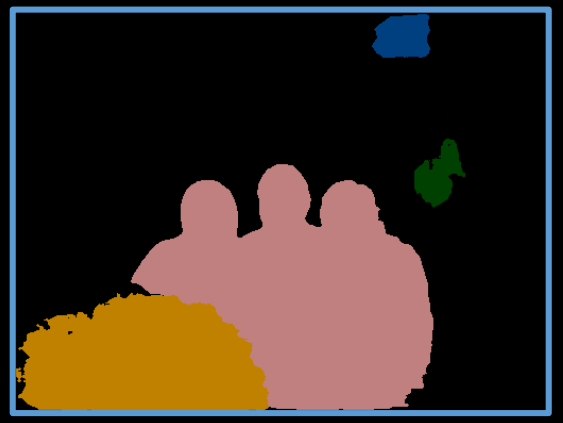
Ground Truth



Unary Term



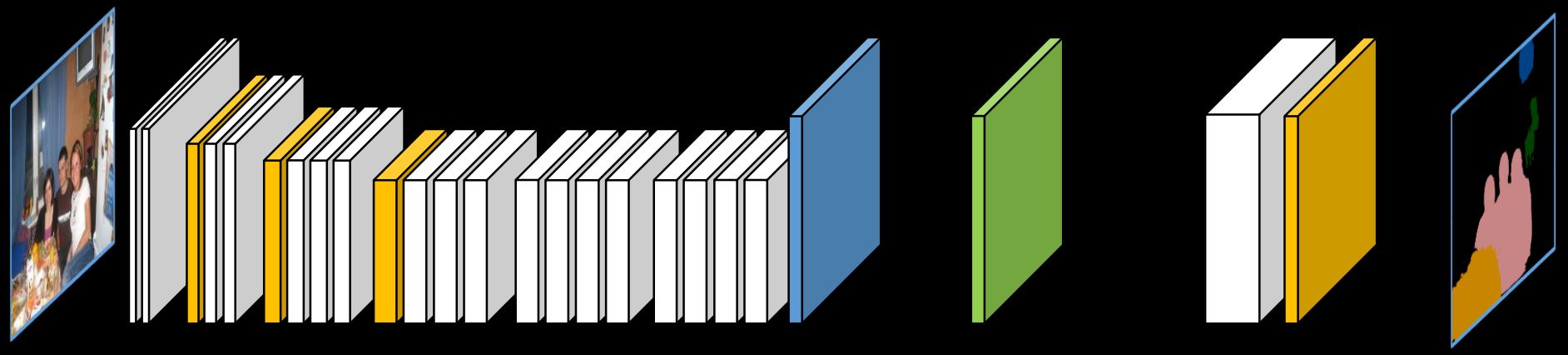
Triple Penalty



Label Contexts

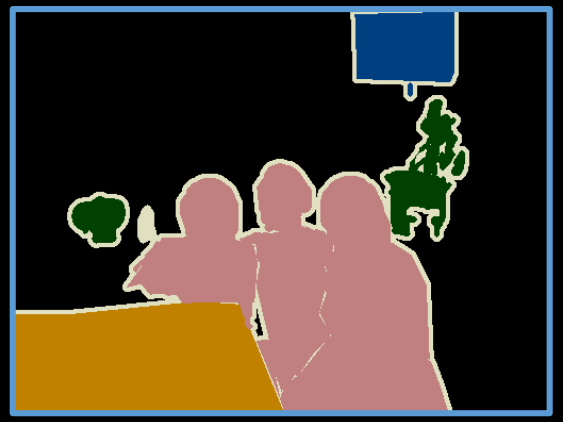
Deep Parsing Network

Joint Tuning

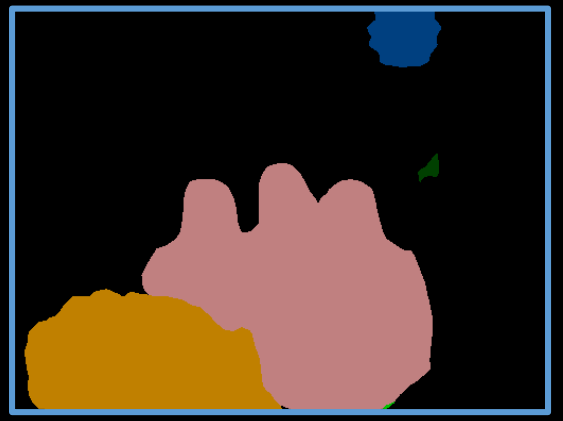


Deep Parsing Network

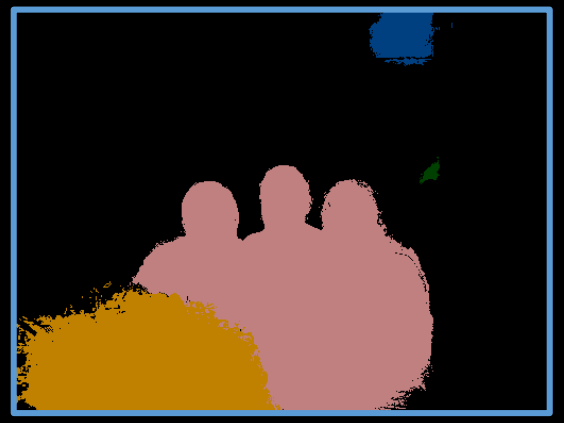
Original Image



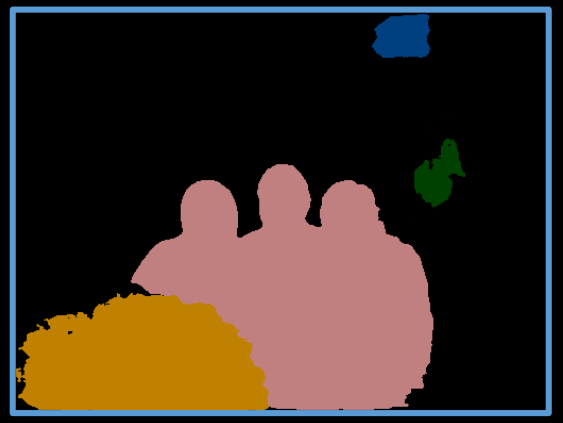
Ground Truth



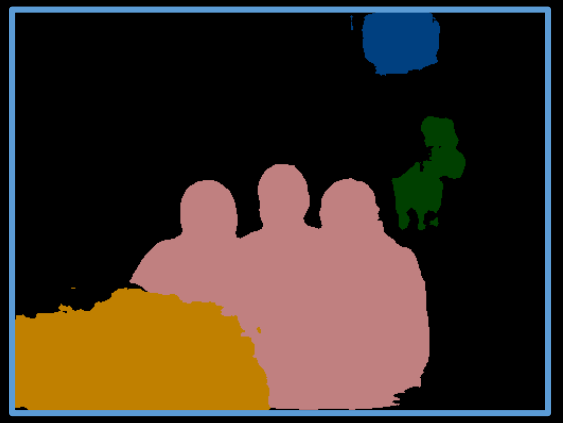
Unary Term



Triple Penalty



Label Contexts



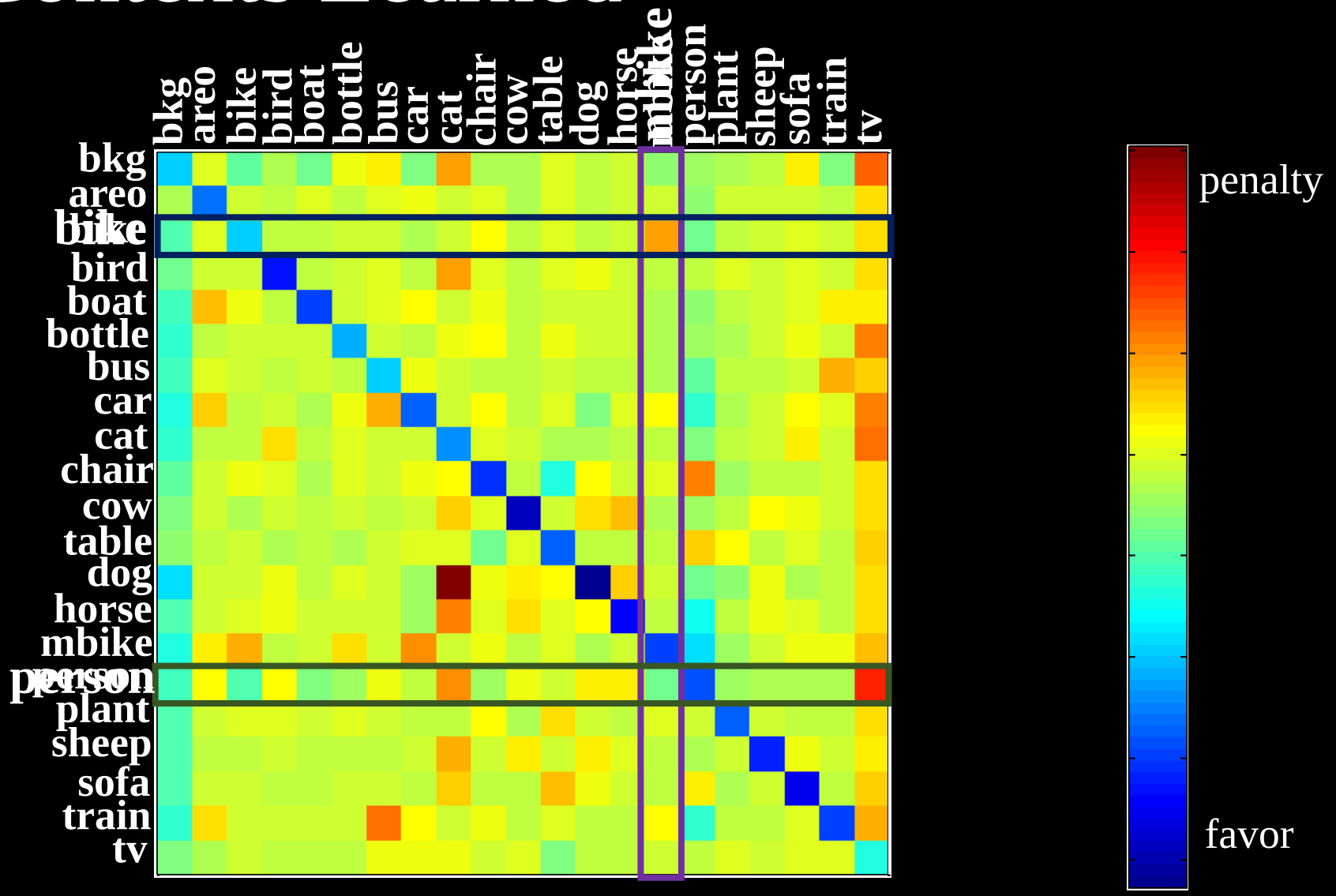
Joint Tuning

Overall Performance (Published Results)

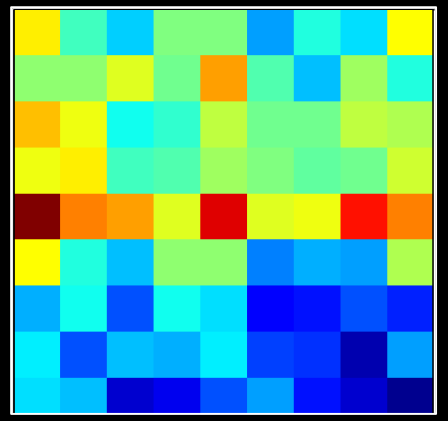
FCN	62.2
DeepLab†	73.9
CRFasRNN†	74.7
BoxSup†	75.2
DPN†	**77.5**

(PASCAL VOC 2012 Challenge test set)

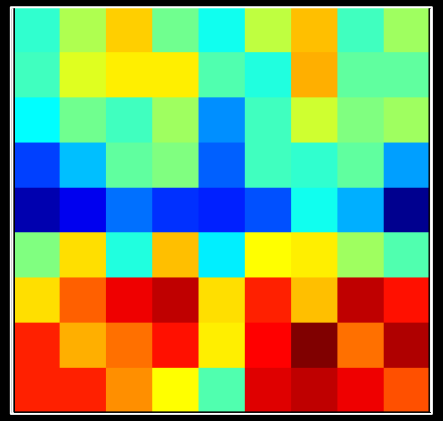
Label Contexts Learned



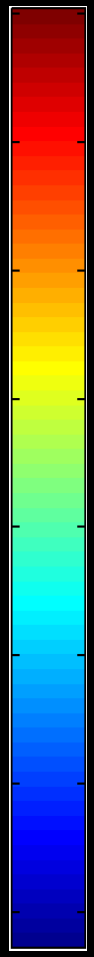
Label Contexts Learned



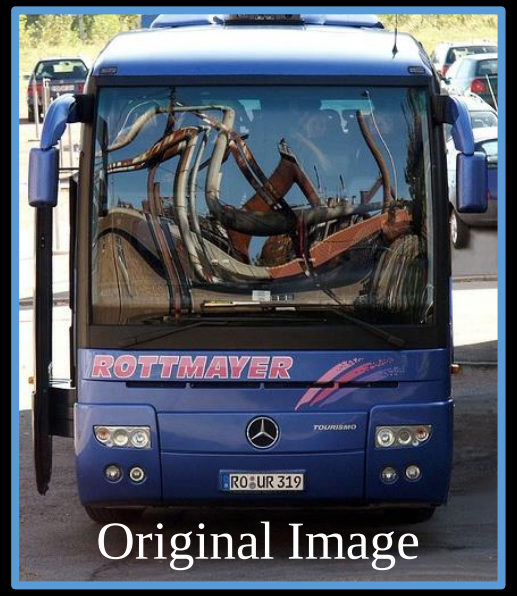
person : mbike

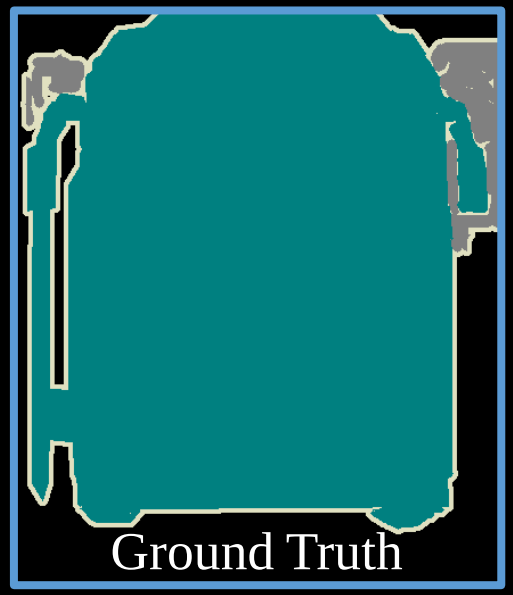


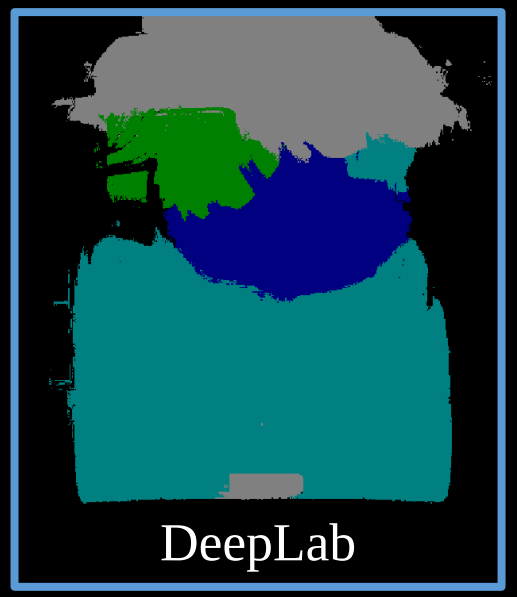
chair : person

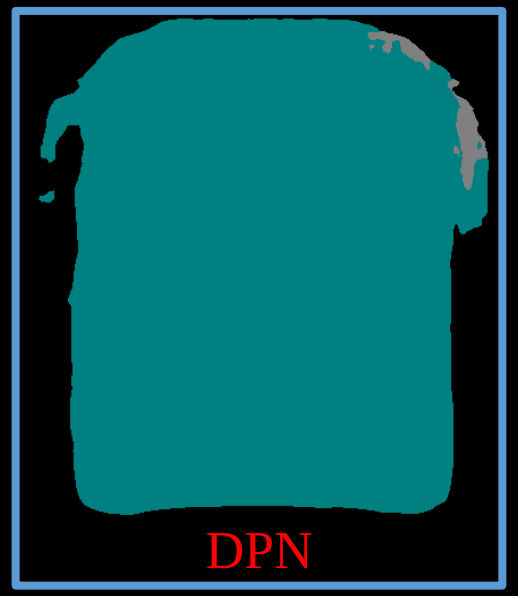


Challenging Case









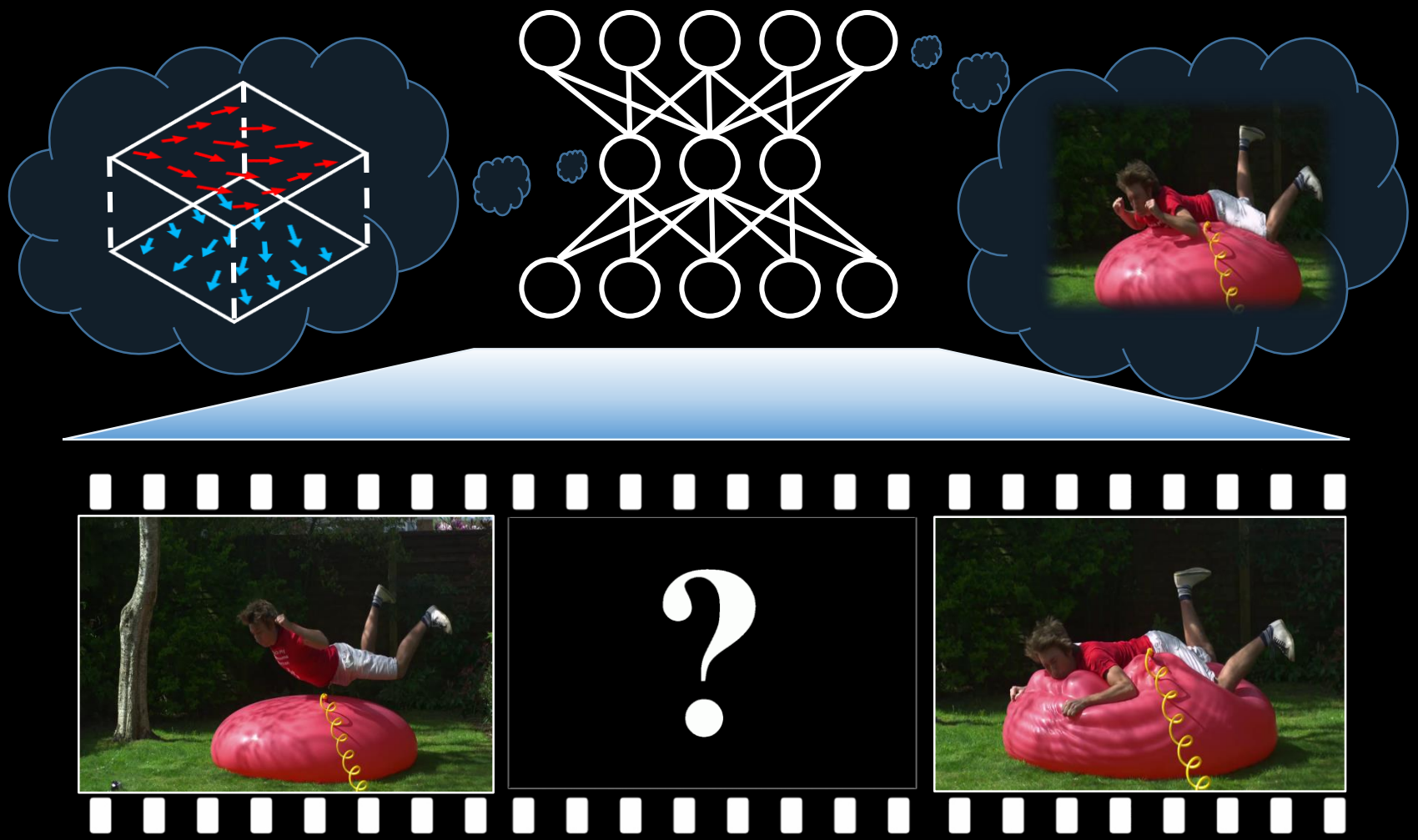
Demo

Part IV: Deep Motion Understanding

Video Frame Synthesis

- Problem

Video interpolation/ extrapolation



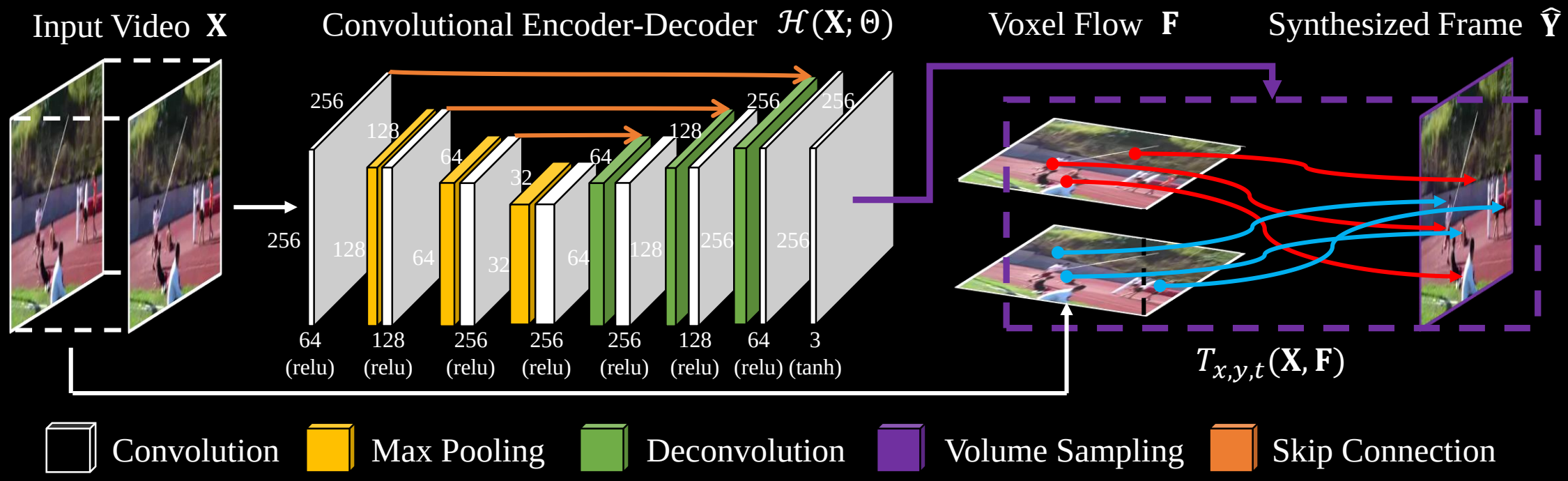
Video Frame Synthesis

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

Deep Voxel Flow

- Motivation

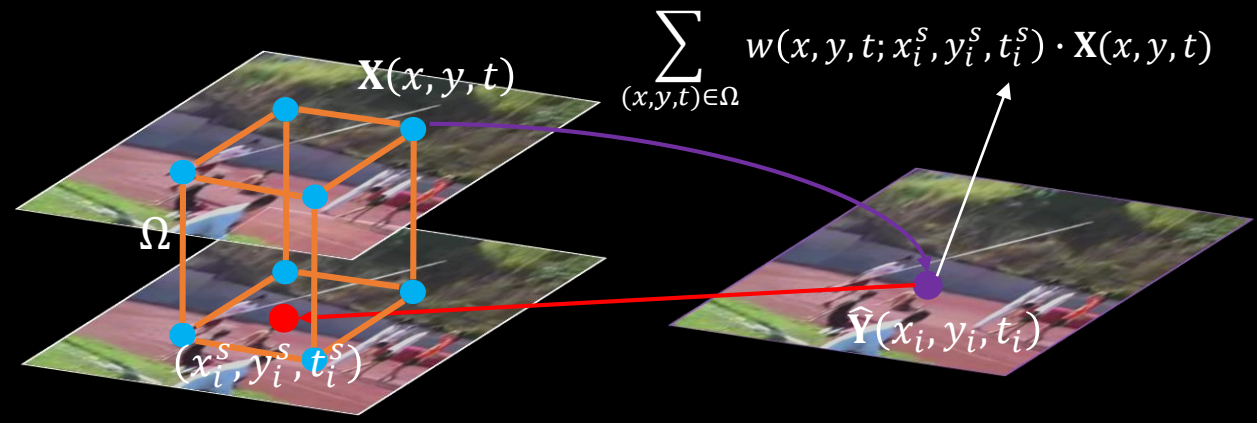
Combining the strength of flow-based and NN-based methods



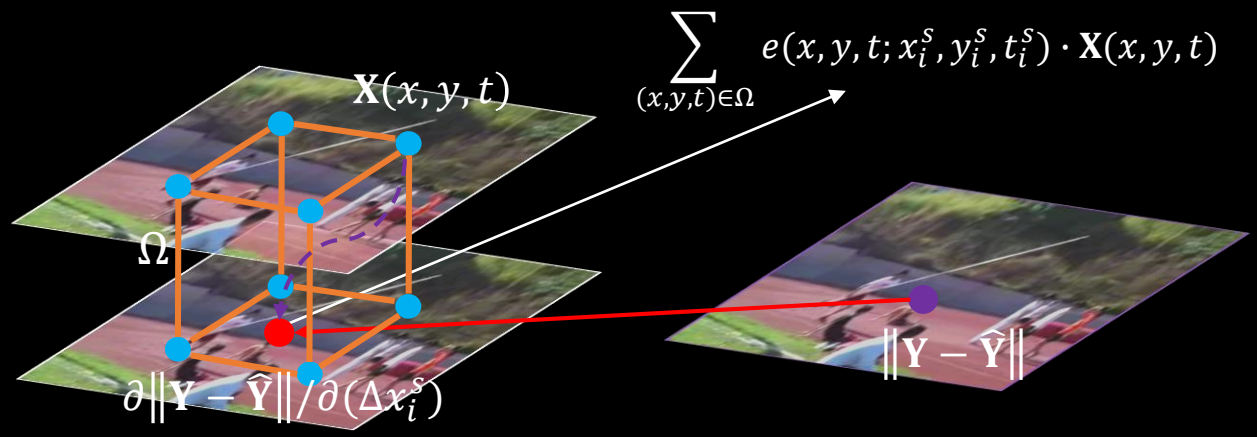
Deep Voxel Flow

- Mechanism

Differentiable spatio-temporal sampling



(a) Forward Pass

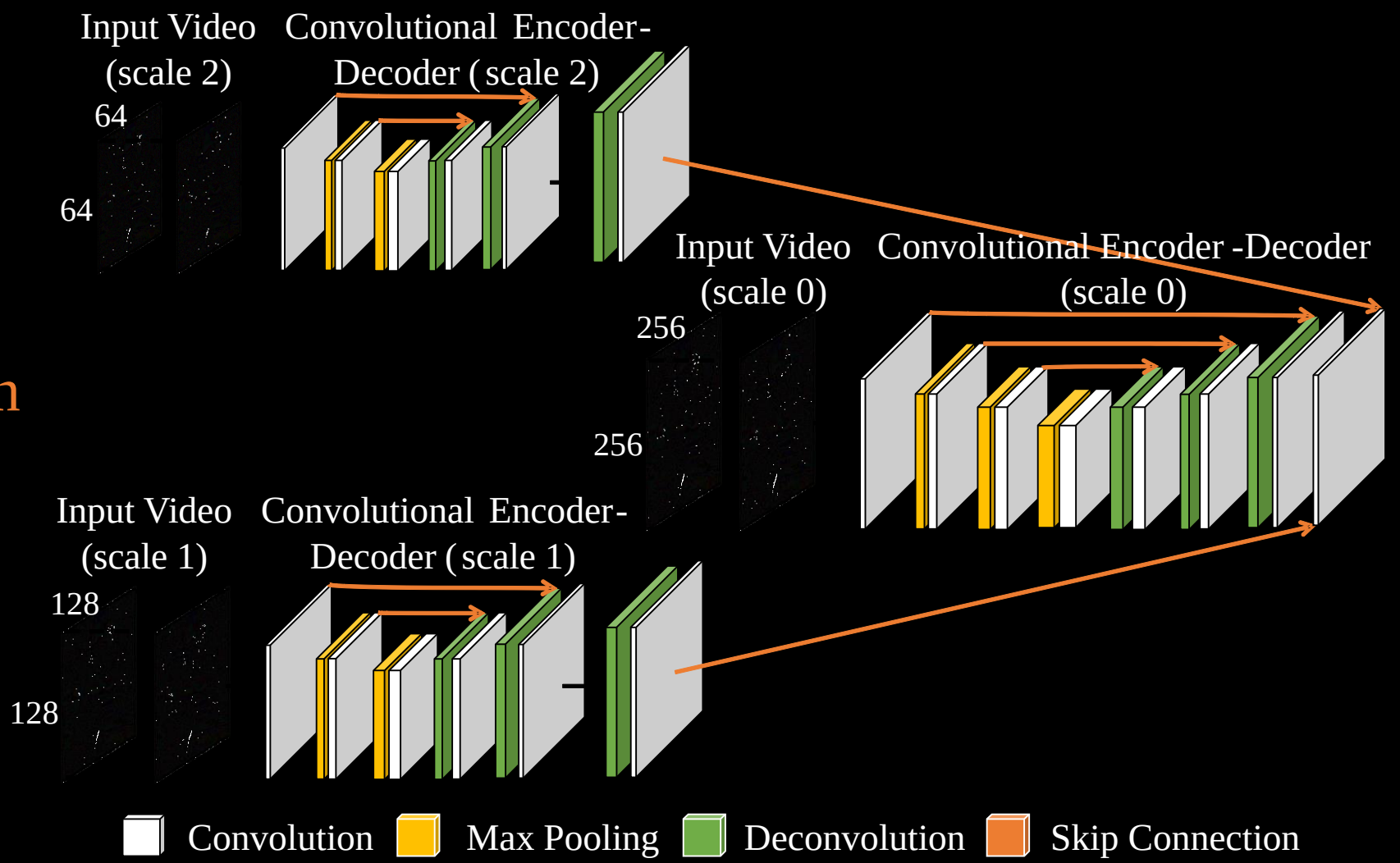


(b) Backward Pass

Multi-scale Deep Voxel Flow

- Pipeline

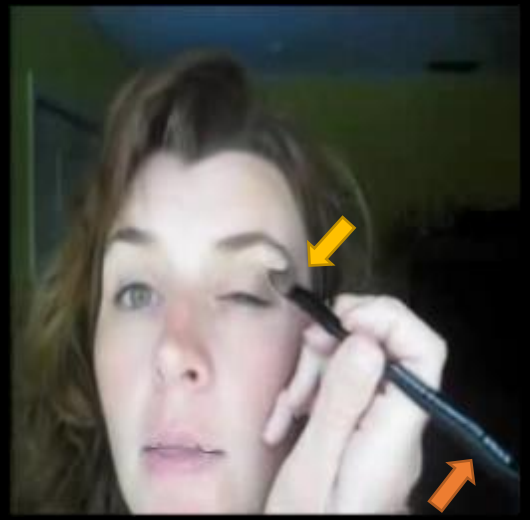
Handle large motion



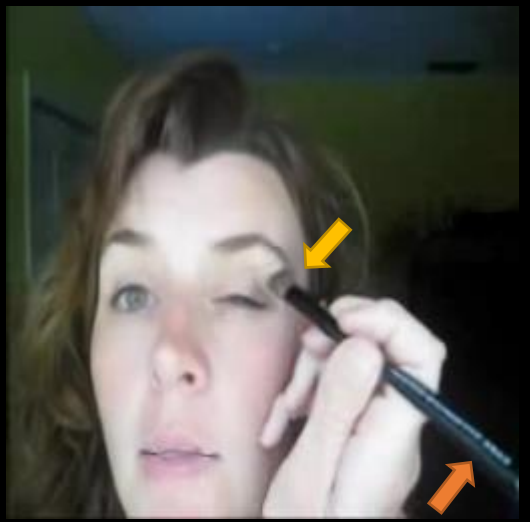
Multi-scale Voxel Flow

- Advantages

(a) 2D Flow + Mask

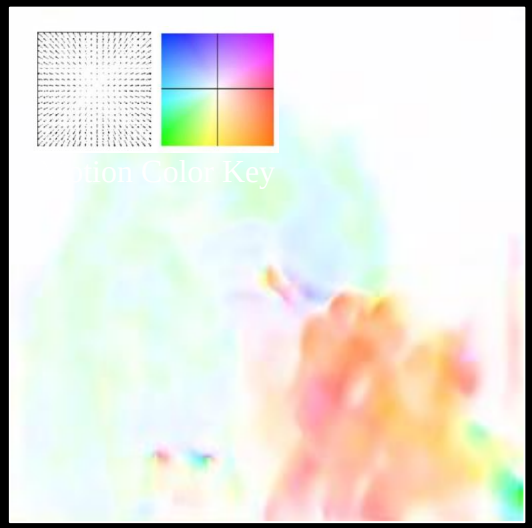


(b) Voxel Flow

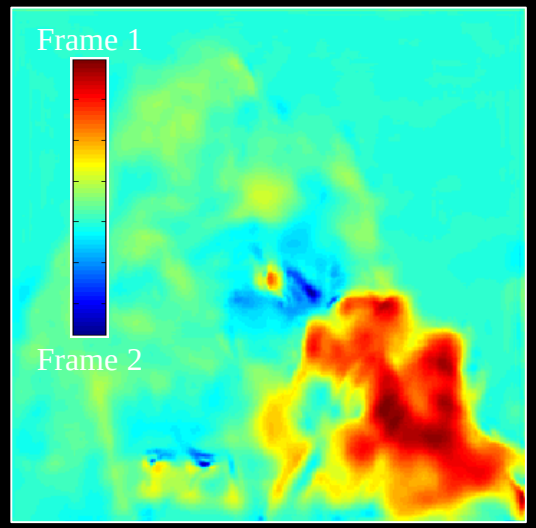


(c) Multi-scale Voxel Flow

(d) Difference Image



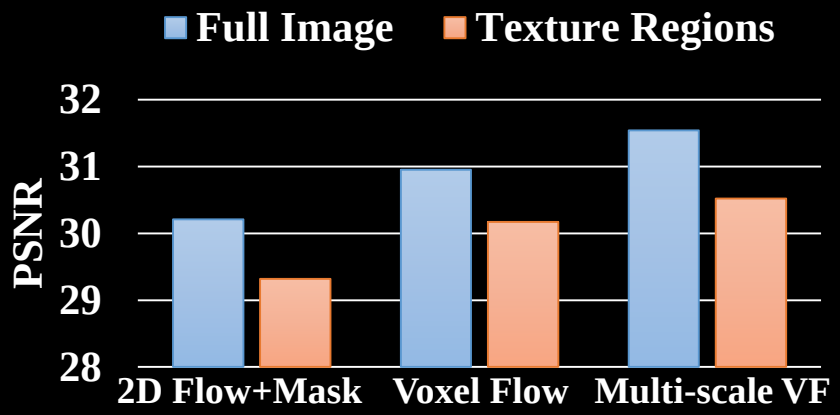
(e) Projected Motion Field



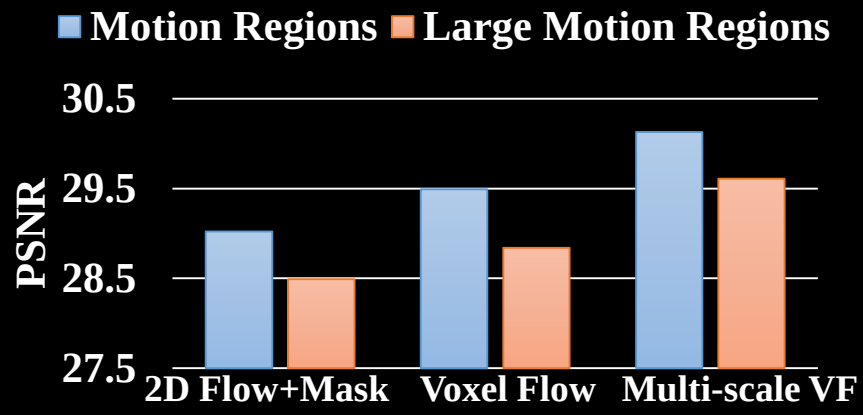
(f) Projected Selection Mask

Multi-scale Voxel Flow

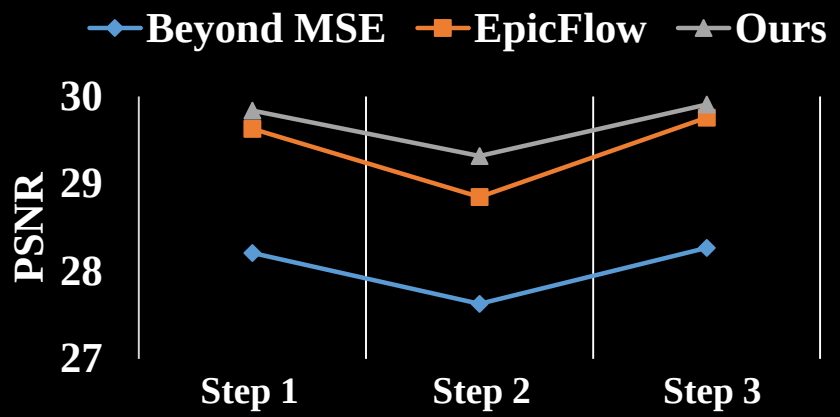
- Ablation Study



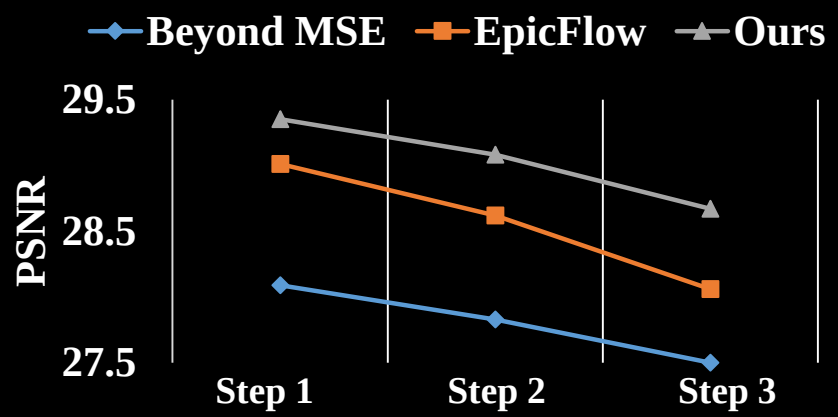
(a) Appearance



(b) Motion



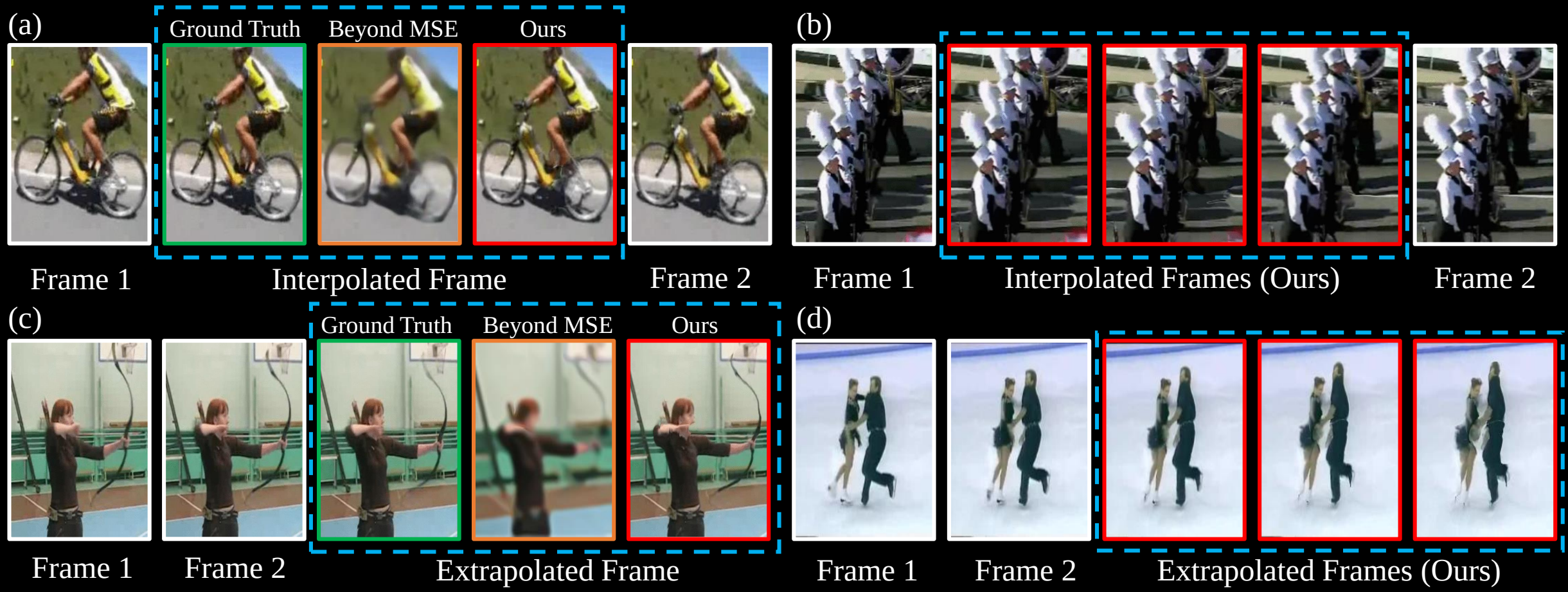
(c) Interpolation



(d) Extrapolation

Comparisons

- UCF-101



Comparisons

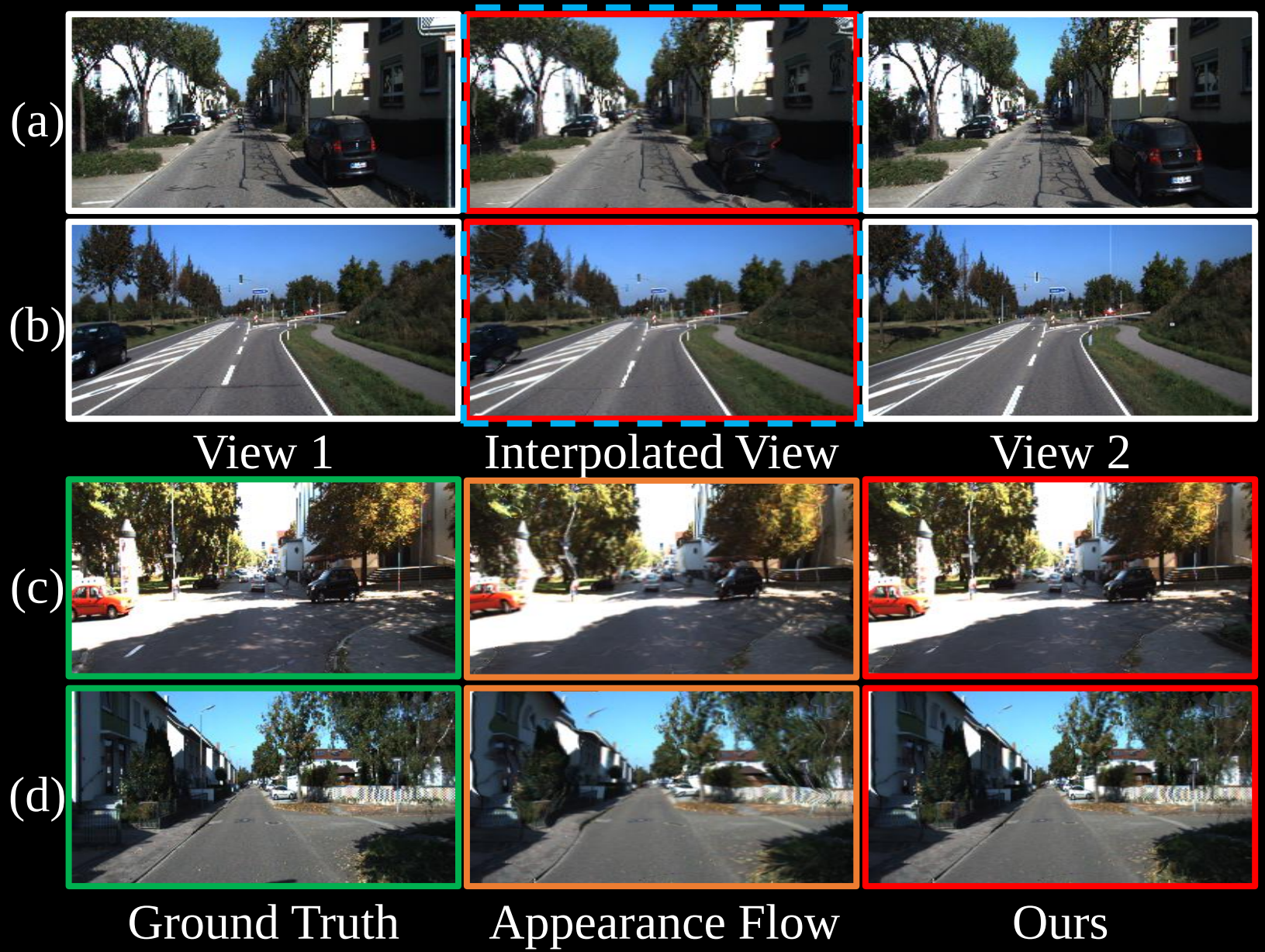
- UCF-101

Comparisons

- UCF-101

Comparisons

- KITTI



Comparisons

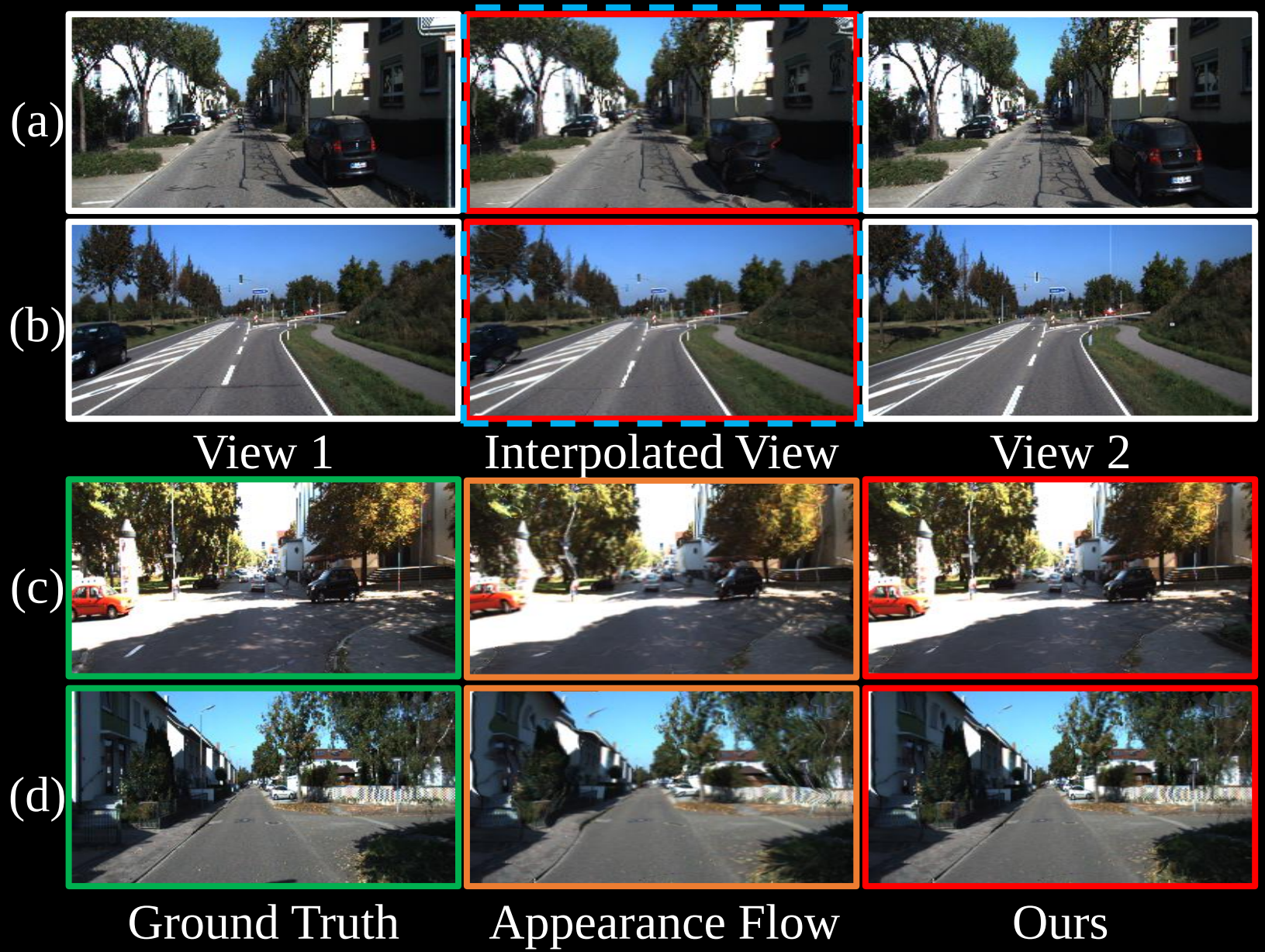
- KITTI

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**

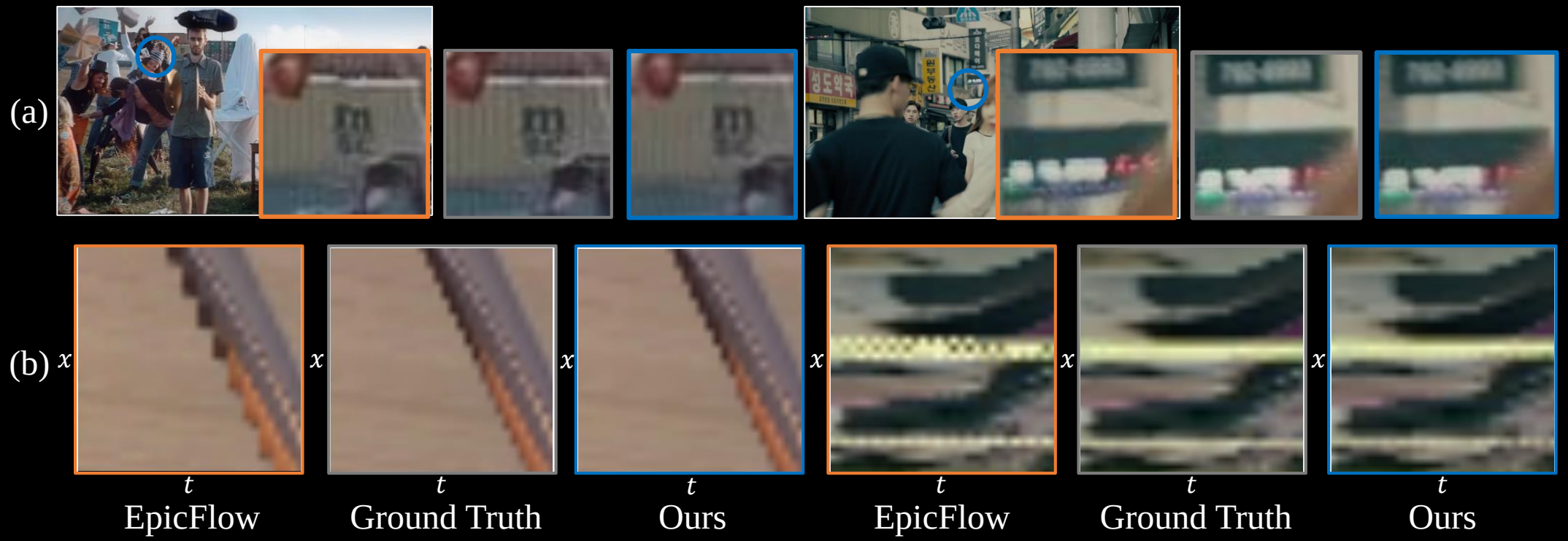
Flow estimation

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

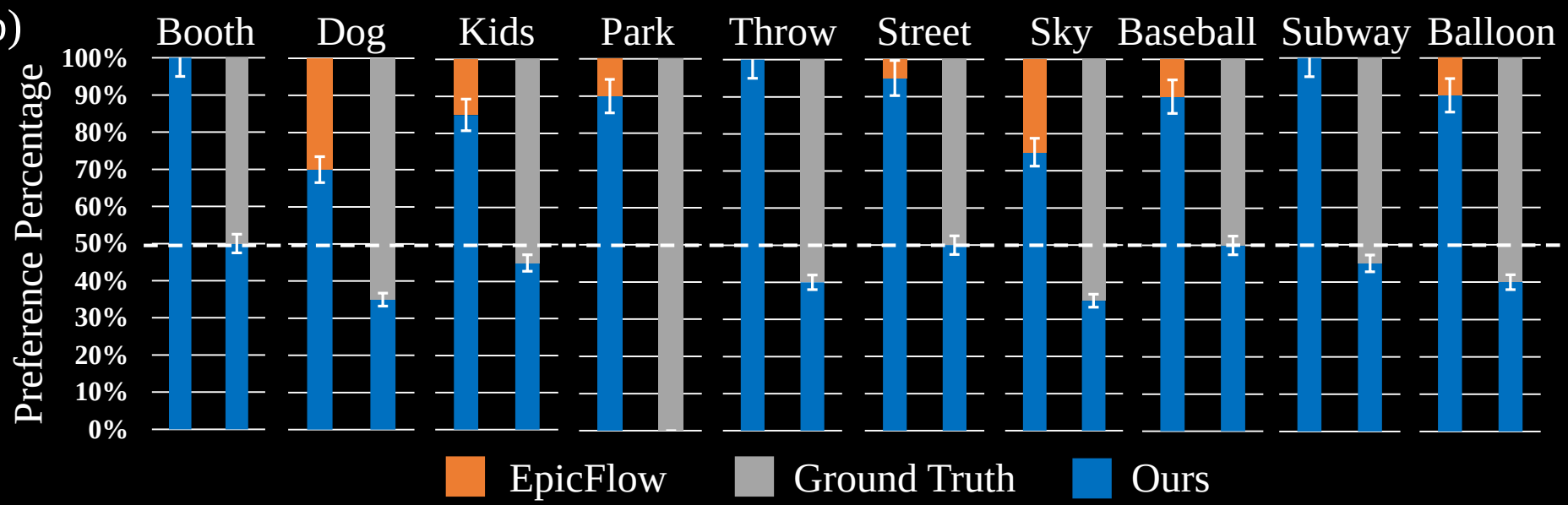
- User Study

(a)

Diagonal-split Comparison

Method 1 \ Method 2

(b)



Real-life Applications

- Demo

Conclusions

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

Achievements (I)

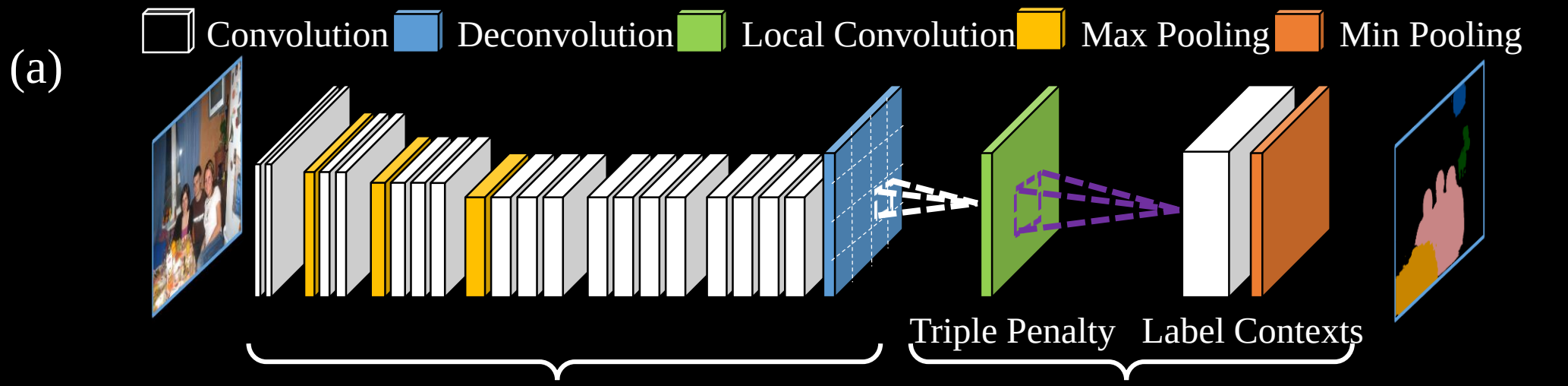
Databases

CelebA

DeepFashion

Achievements (II)

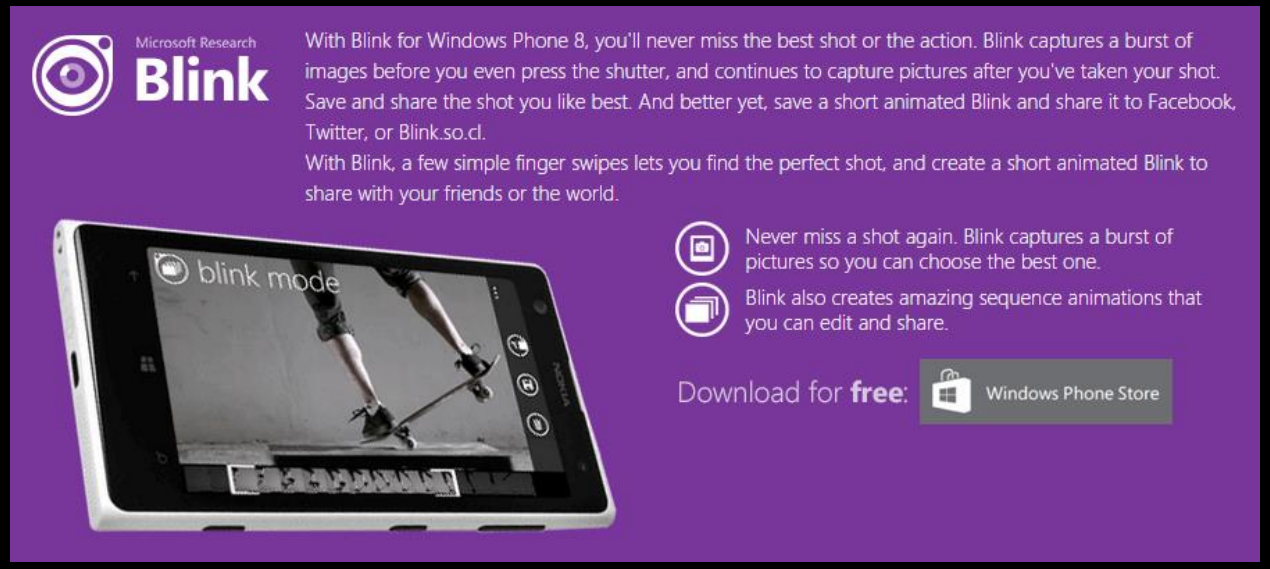
Theory



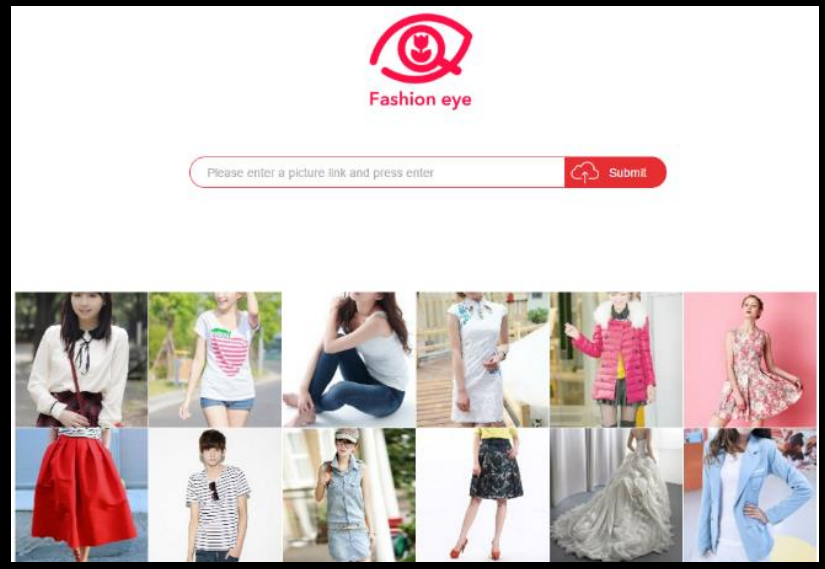
Incorporate Various Structures into Deep Learning

Achievements (III)

Products



Microsoft Blink



SenseTime FashionEye

Future Work

- From “single-level structure” to “multi-level structure”
- From “passive deep learning” to “active deep learning”
- From “one-pass inference” to “interactive reasoning”

Collaborators

Xiaoxiao Li

Sijie Yan

Shi Qiu

Ping Luo

Chen Change Loy

Xiaogang Wang

Xiaoou Tang

Thanks!

If I have been able to see further, it was only because I stood on the shoulders of giants.

Homepage: *http://personal.ie.cuhk.edu.hk/~lz013/*