Learning and Vision Group, NUS (NUS-LV)

Deep Learning and Biometrics

Shuicheng YAN
eleyans@nus.edu.sg
National University of Singapore

[Special thanks to Min LIN, Qiang CHEN, Luoqi LIU, Xiaodan LIANG, Si LIU, Xiaobo SHU, and Zhiheng NIU]

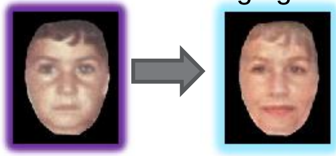
Learning and Vision Research Group (LV)

- Founded early 2008, frequently 20-30 members
- Focus on multimedia, computer vision and machine learning

Interests on Hard/Soft Biometrics in NUS-LV

Beauty e-Experts

Personalized Aging

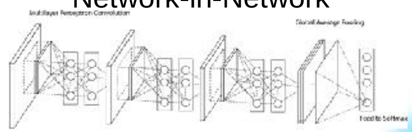


How old? Gender?

How to beautify?

What shall she look like after N years?

Network-in-Network



Human Parsing

Whom is she/he?

Whole body details?
(hair, clothes, bag, etc.)

Person Re-ID?

……

Face

Body

NUS National University of Singapore

I. Biometrics without Deep Learning

(Beautification/De-aging vs. Aging)

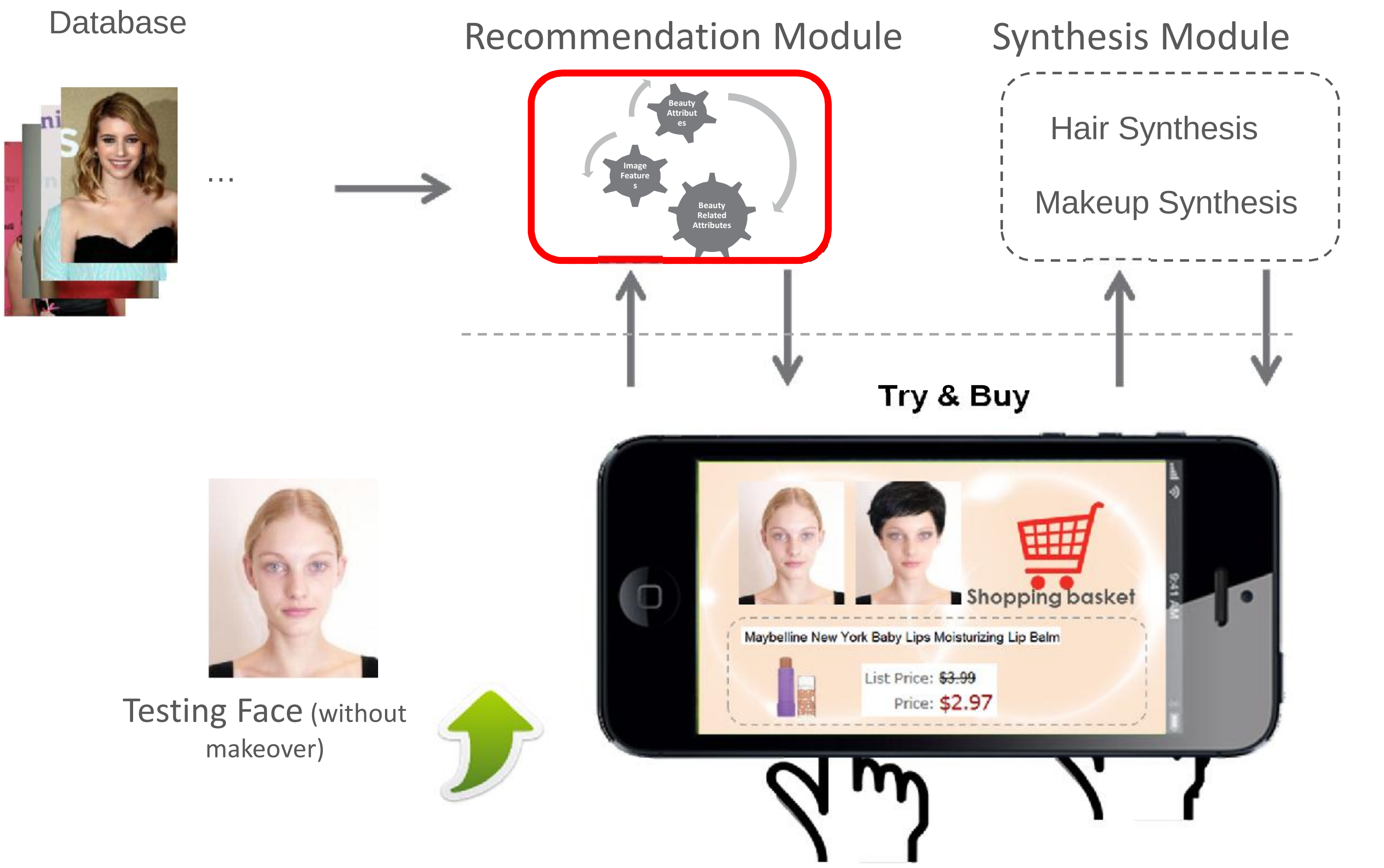
Task I: Face Beautification/De-aging

(Beauty e-Experts)

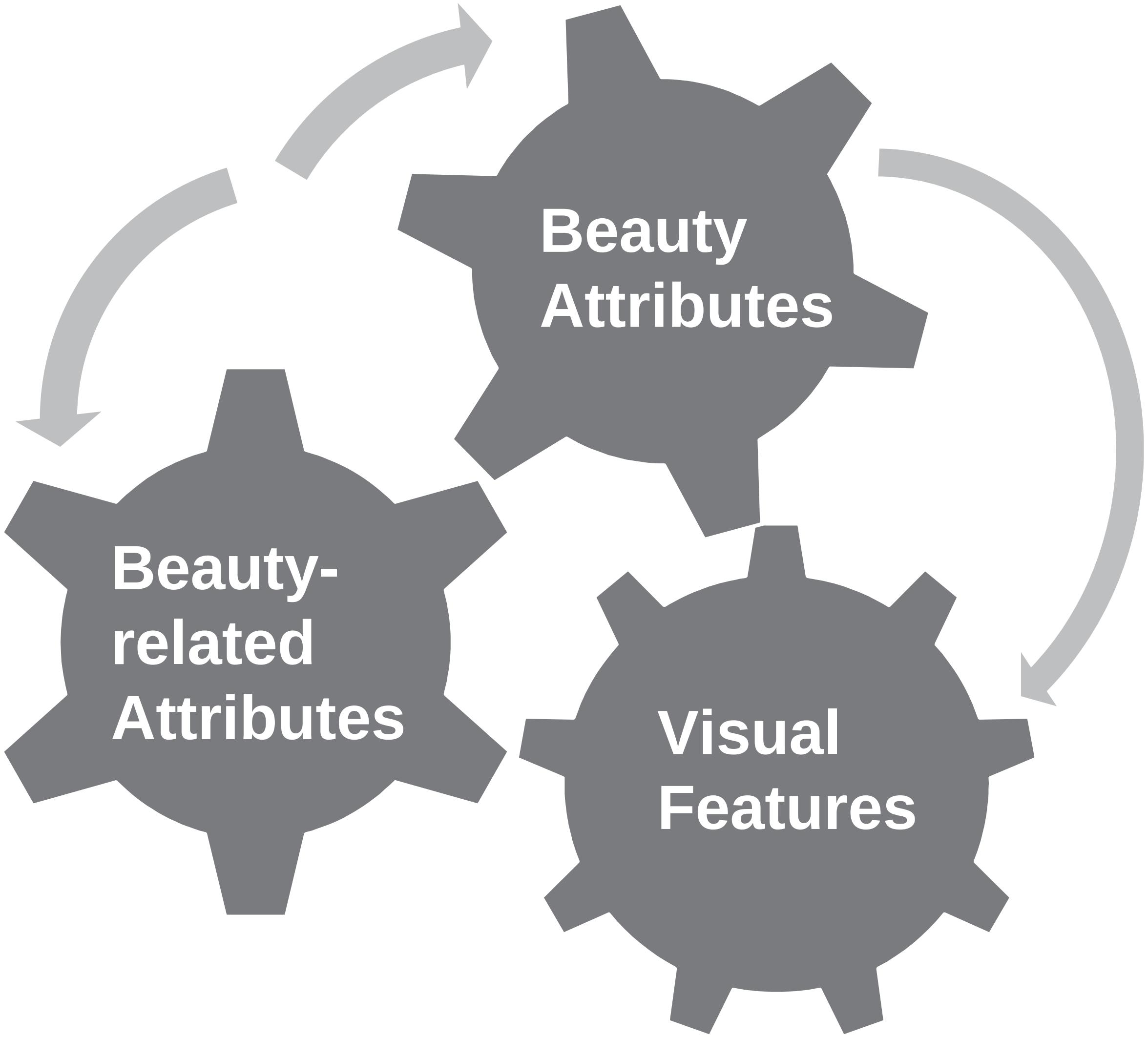
Makeover (makeup+ hairstyle) Process

- The makeover process

System Flowchart



Recommendation Module

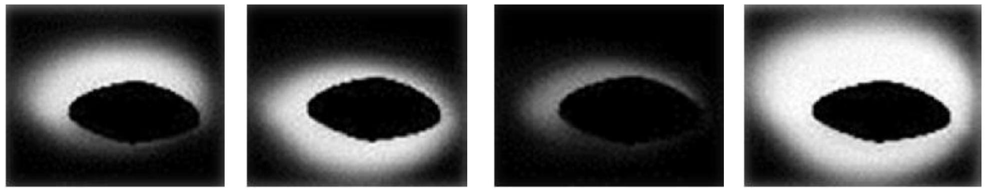


Beauty Attributes

Hair length
Hair color
Hair shape
Hair bangs
Hair volume

Spectral matting

Eye shadow shape



Spectral matting+ clustering

Eye shadow color

Clustering

Foundation

Clustering

Lip gloss

Clustering

Totally, we define 9 kinds of beauty attributes(directly related with real cosmetic products).

Beauty-related Attributes

Face shape

long

oval

round

Lip thickness

thick

normal

Ocular distance

wide

normal

narrow

Race

western

eastern

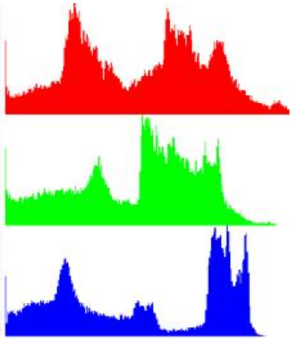
⋮

Totally, we define 21 kinds of beauty-related attributes:

(1) Unchanged during makeover process

(2) Strong correlations with beauty attributes

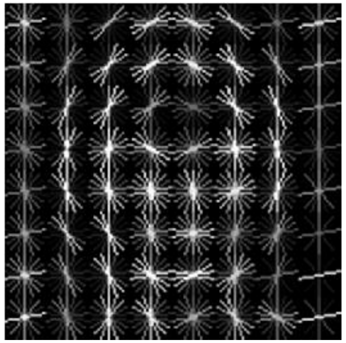
Visual Features



$$\text{Mean}: \quad E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij}$$

$$\text{StandardDeviation}: \quad \sigma_i = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^2}$$

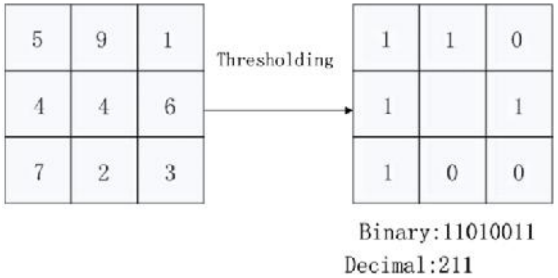
$$\text{Skewness}: \quad s_i = \sqrt[3]{\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^3}$$

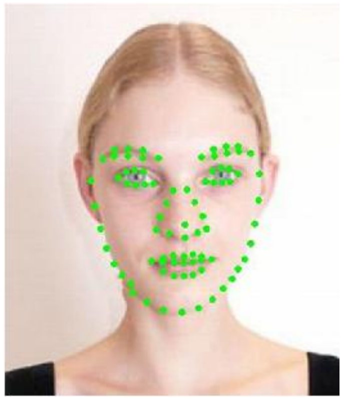


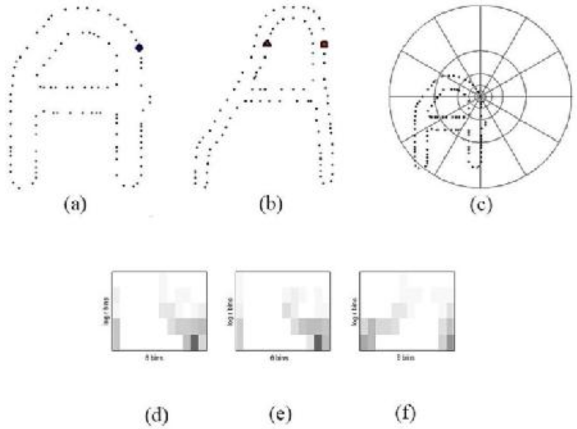
Color Histograms

Color Moments

Histogram of Gradients





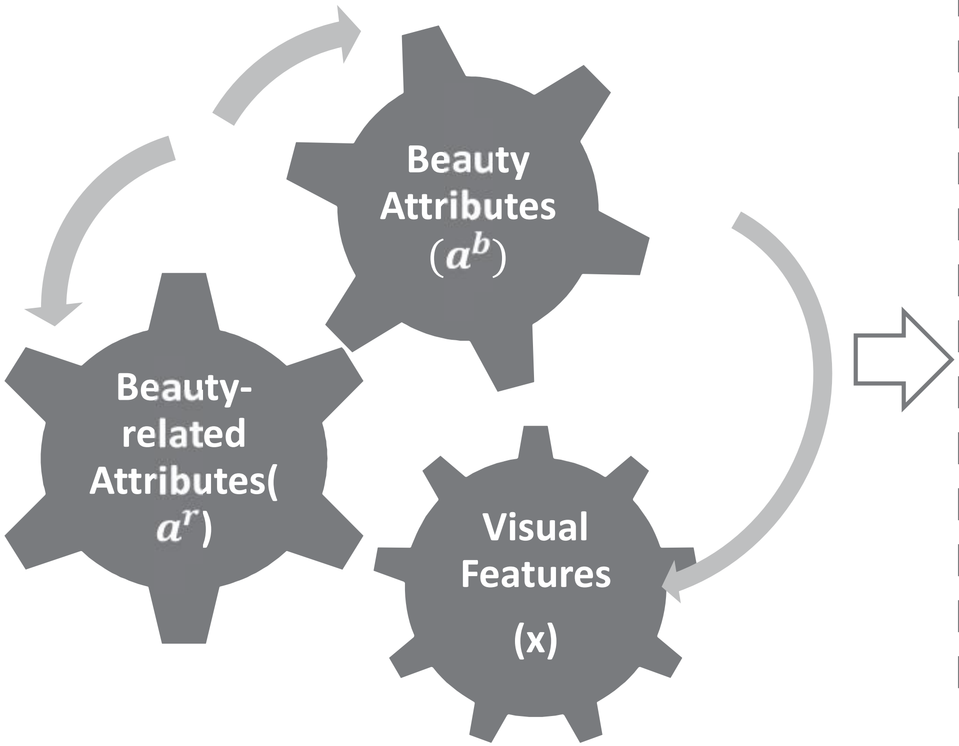


Local Binary Patterns

ASM Parameters

Shape Context

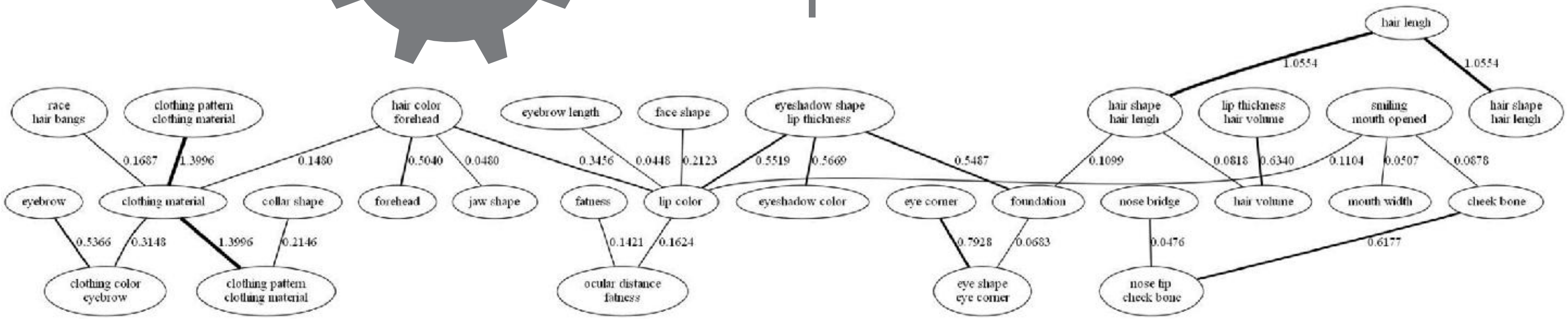
Recommendation Model: Formulation



Gibbs Distribution

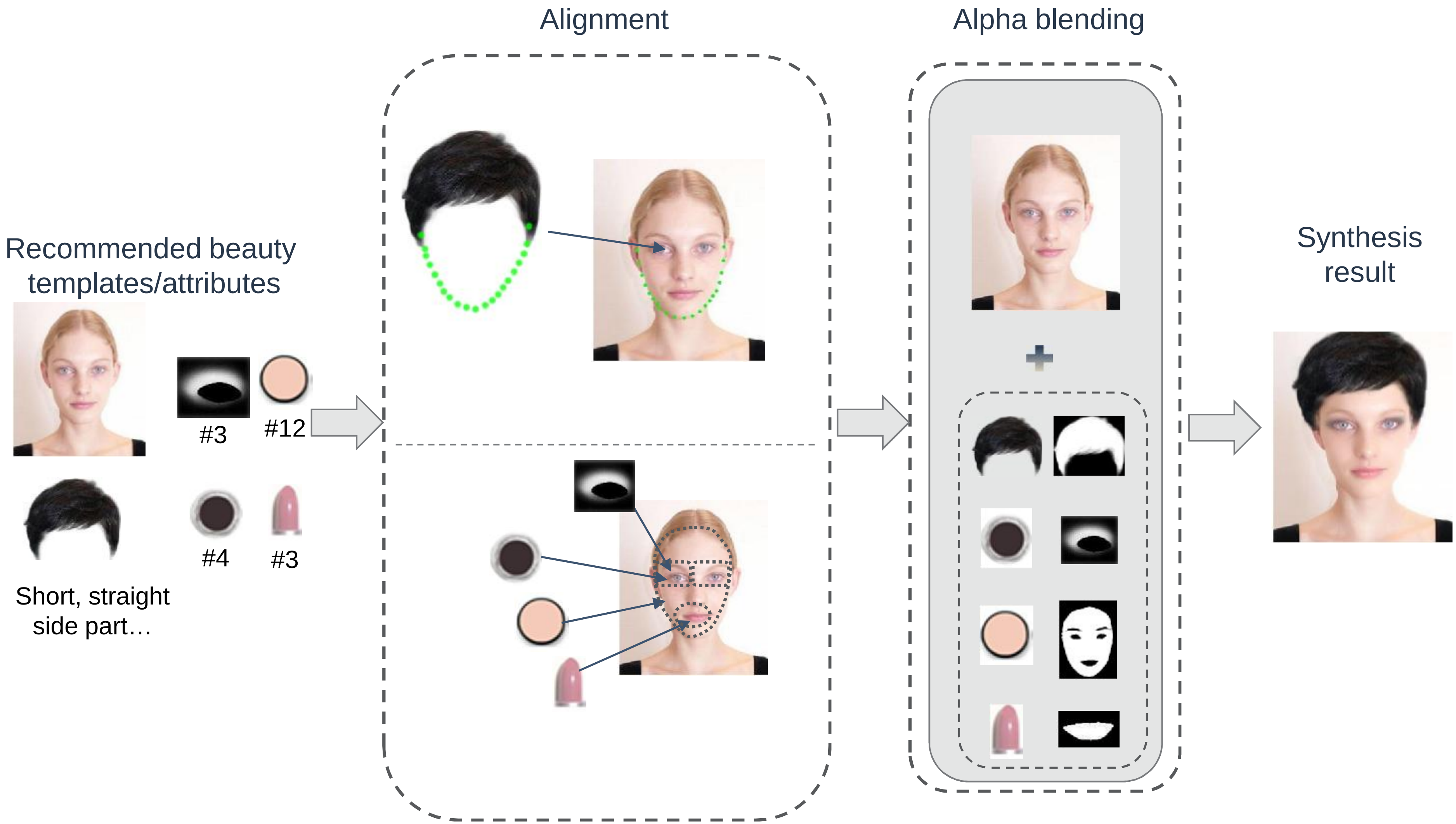
$$p(\mathbf{a}^b, \mathbf{a}^r \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(-E(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x})\right)$$

$$Z(\mathbf{x}) = \sum_{\mathbf{a}^b, \mathbf{a}^r} \exp\left(-E(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x})\right)$$



Super-graph [super-vertex and Maximum Spanning Tree] to define ***E***

Synthesis Module



Exemplar Synthesis Process

Foundation

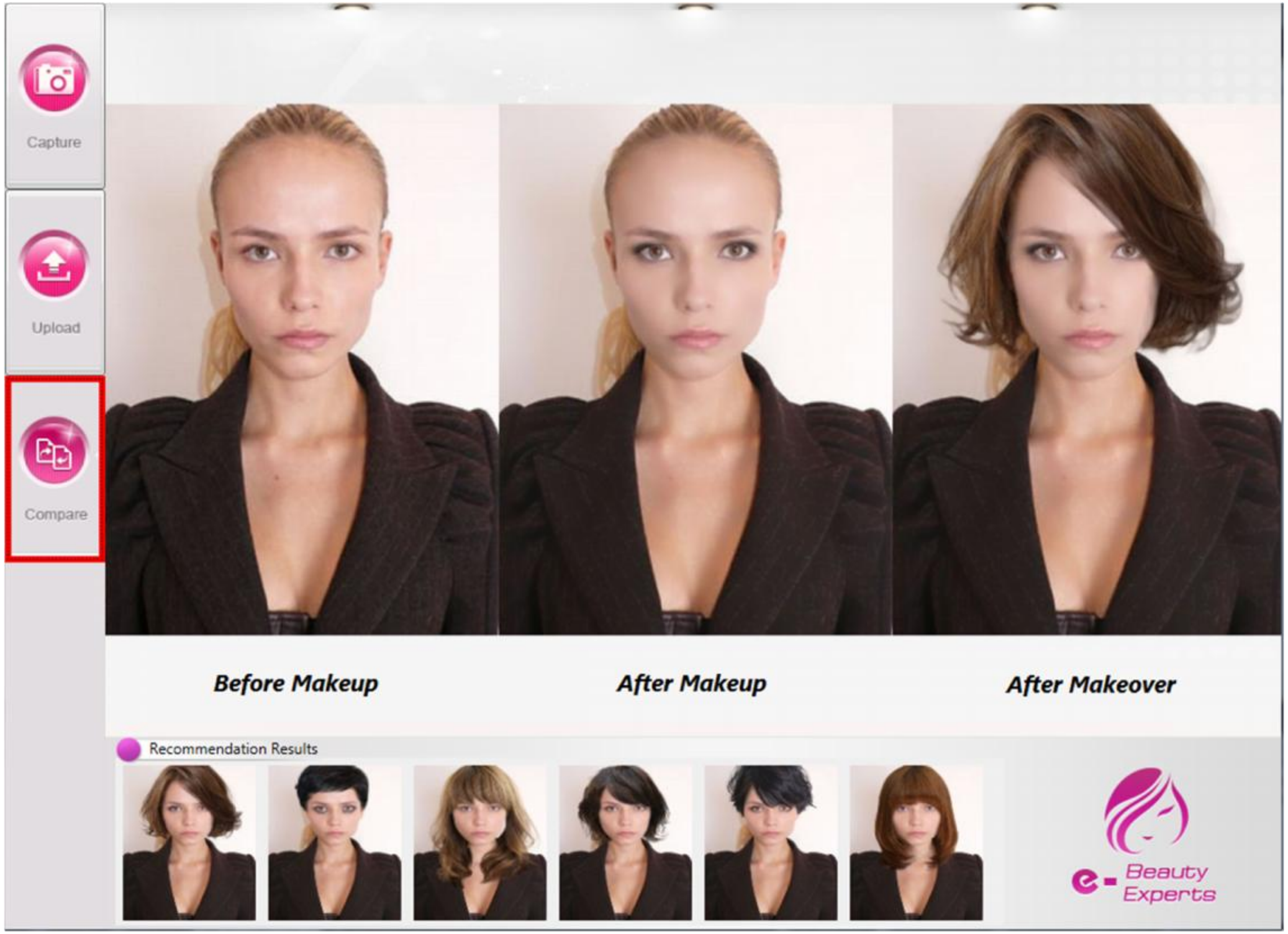
Lip Gloss

Eye Shadow

Hair style

Recommendation and Synthesis Results

Recommendation and Synthesis Results



Task II: Face Aging Progression

(Personalized Aging)

Personalized Age Progression

- Aim to render aging faces in a personalized way
- Personalized aging face contains the aging layer (e.g. wrinkles) and the personalized layer (e.g. mole, unchanged)

Invariant, shared among different age groups

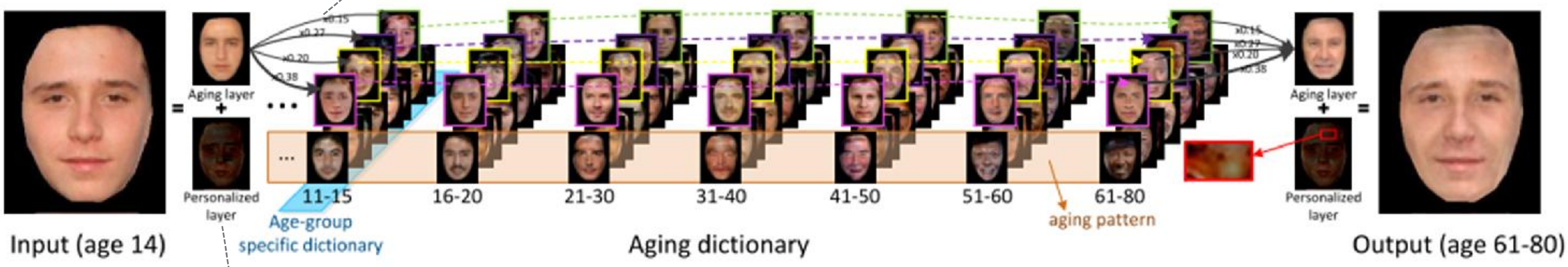


Illustration of Framework

Unchanged in aging progression

Aging Dictionary Learning with Neighbor-group Pairs

- Couple-aware aging dictionary learning

$$\min_{\mathbf{D},\mathbf{A},\mathbf{P}} \sum_{g=1}^{G-1} \Big\{ \|\mathbf{X}^g - \boxed{\mathbf{H}^g\mathbf{D}^g}\mathbf{A}^g - \mathbf{P}^g\|_F^2 + \gamma\|\mathbf{P}^g\|_F^2$$

$$+ \left\|\mathbf{Y}^g - \boxed{\mathbf{H}^{g+1}\mathbf{D}^{g+1}}\mathbf{A}^g - \mathbf{P}^g\right\|_F^2 + \lambda\|\mathbf{A}^g\|_1 \Big\}$$

$$\text{s.t.} \quad \|\mathbf{D}^g(:,d)\|_2 \le 1, \forall d \in \{1,\cdots,k\}, \forall g \in \{1,\cdots,G\}$$

Aging dictionary

- Bi-level aging dictionary learning

$$\min_{\mathbf{D}^g,\mathbf{D}^{g+1}} \|\mathbf{X}^g - \mathbf{H}^g\mathbf{D}^g\mathbf{A}^g - \mathbf{P}^g\|_F^2 + \left\|\mathbf{Y}^g - \mathbf{H}^{g+1}\mathbf{D}^{g+1}\mathbf{A}^g - \mathbf{P}^g\right\|_F^2$$

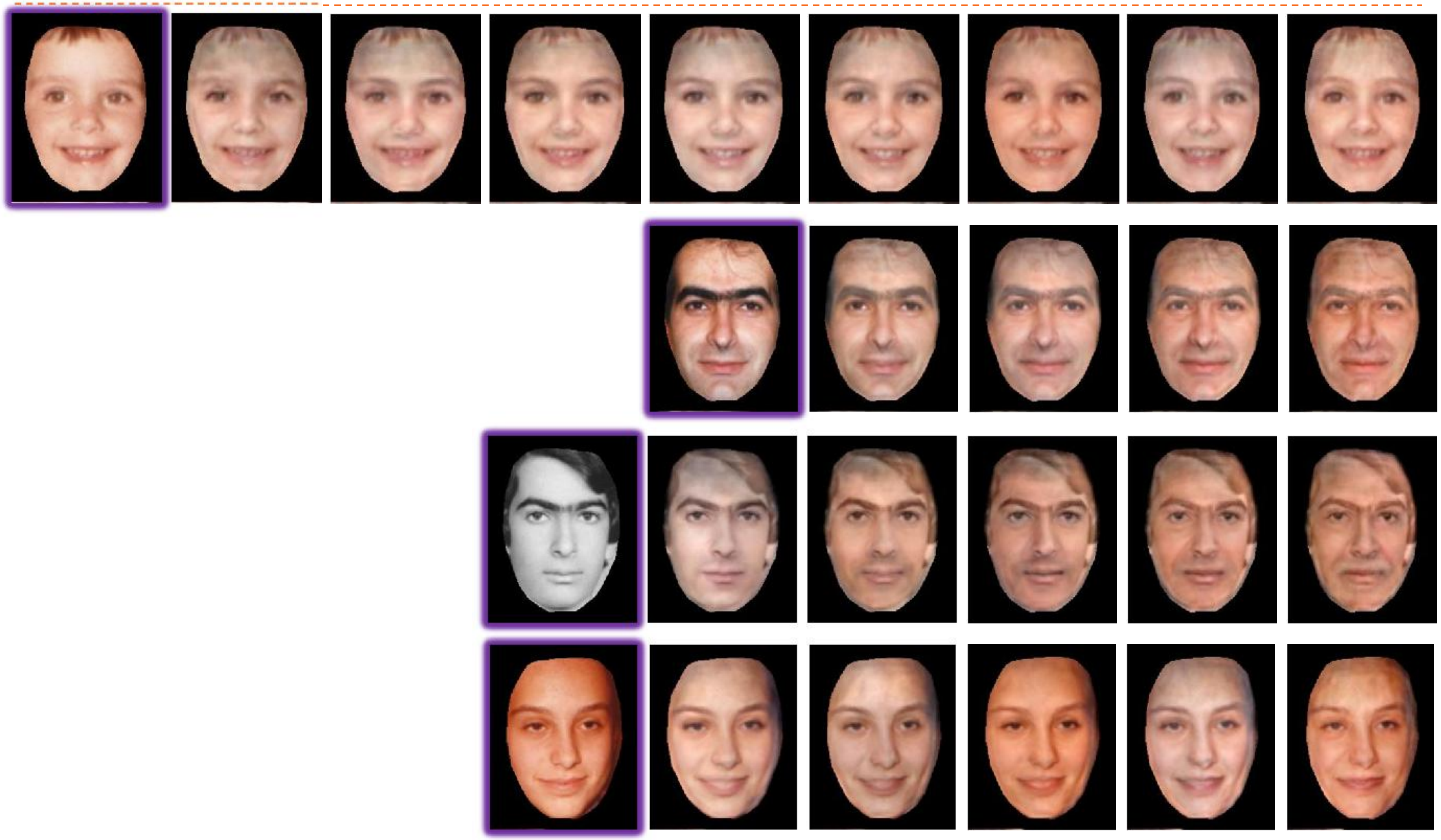
$$s.t. \quad \mathbf{A}^g = \arg\min_{\mathbf{Z}^g} \|\mathbf{X}^g - \mathbf{H}^g\mathbf{D}^g\mathbf{Z}^g - \mathbf{P}^g\|_F^2 + \lambda_1\|\mathbf{Z}^g\|_1 + \lambda_2\|\mathbf{Z}^g\|_F^2$$

$$\mathbf{P}^g = \arg\min_{\mathbf{Q}^g} \|\mathbf{X}^g - \mathbf{H}^g\mathbf{D}^g\mathbf{A}^g - \mathbf{Q}^g\|_F^2 + \gamma\|\mathbf{Q}^g\|_1$$

$$\|\mathbf{D}^c(:,l)\|_2 \le 1, \; l = 1,\ldots,k, \; and\, c = \{g, g+1\}.$$

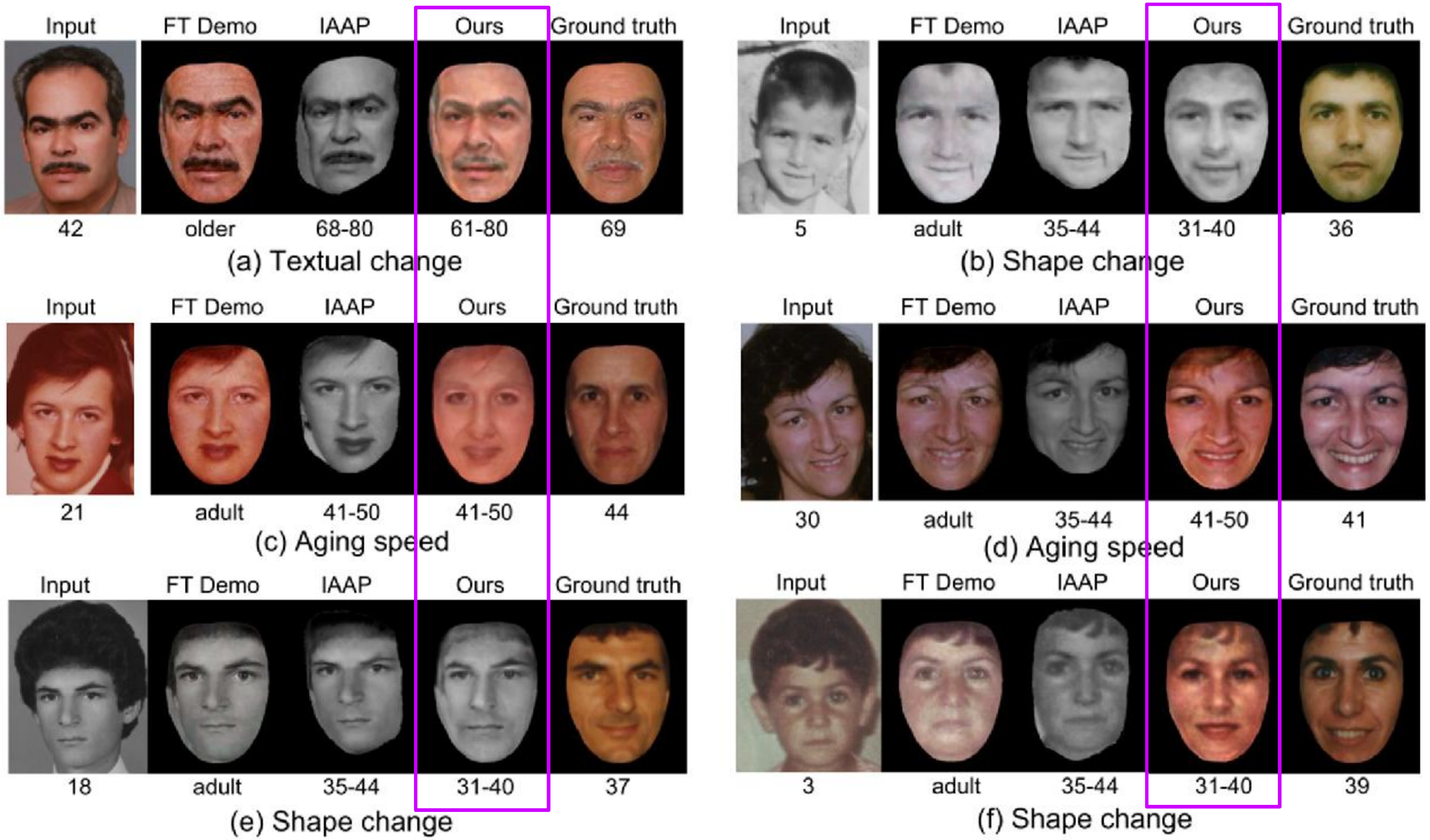
Aging Results

Aging Results



1-5 6-10 11-15 16-20 21-30 31-40 41-50 51-60 61-80

Comparison with Ground Truth



- FT Demo: http://cherry.dcs.aber.ac.uk/Transformer/kinship-aging
- IAAP: I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz. Illumination-aware age progression. In CVPR, 2014.

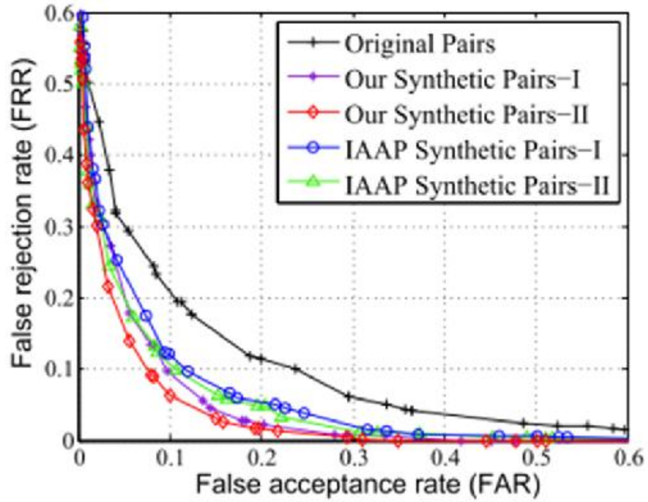
Cross-Age Face Verification by Aging Synthesis

- Face verification with a system with 99.70% accuracy on LFW

--- Our Synthetic Pairs use our aging synthesis method

--- IAAP Synthetic Pairs use our IAAP method

--- "I" and "II" denote using actual age and estimated age, experiments on FG-NET.



Pair settings	Original Pairs	IAAP Synthetic Pairs		**Our Synthetic Pairs**	
		I	II	I	**II**
EER (%)	14.89	10.91	10.36	9.72	**8.53**

II. Biometrics with Deep Learning

Deep Learning Ecosystem in NUS-LV Lab

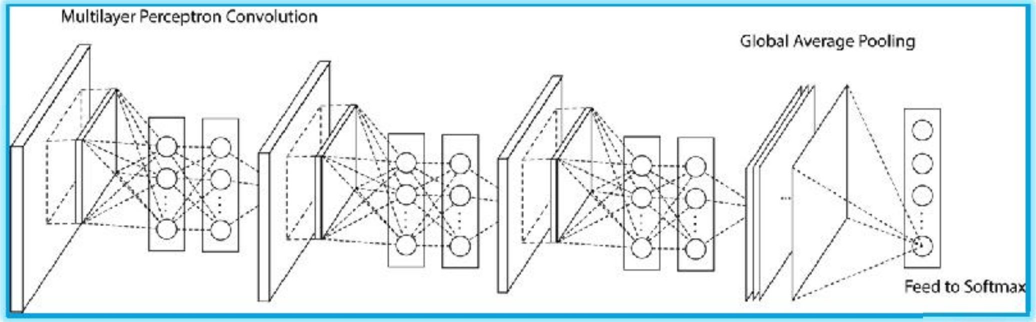
Purine:

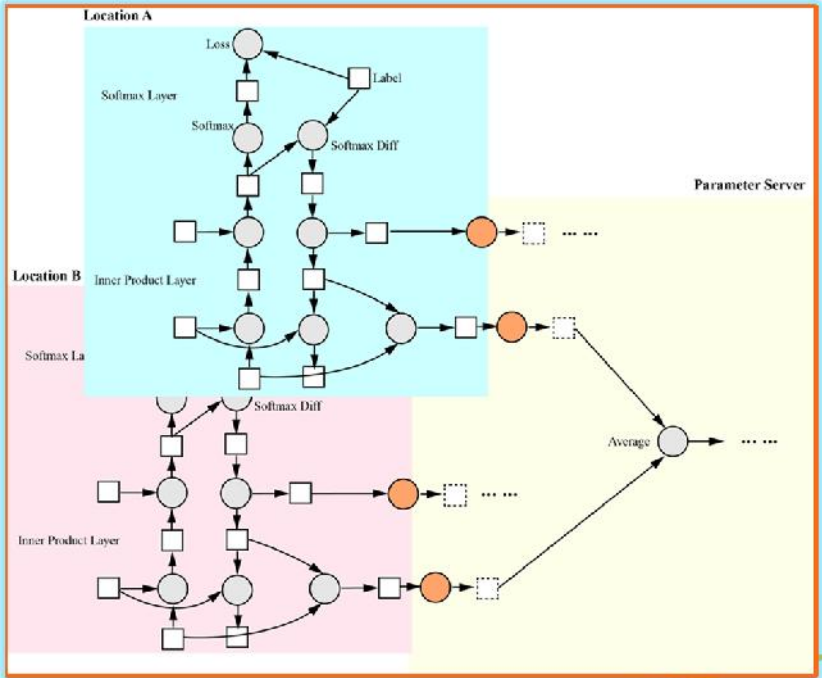
General, bi-graph based DL framework
Multi-PC Multi-CPU/GPU
Linear speedup
High re-usability

Support

Brain-like + Baby-like:

Brain-like network structures and baby-like self/endless learning process



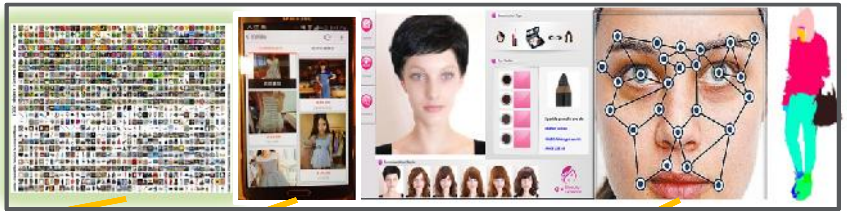


Algorithms

Landing

Visual Perception + Big Data Modeling/Learning

Object/products/human analytics, and other non-visual big data



Architecture

Applications

1. 4 winner awards in VOC
2. One 2nd prize in VOC
3. 2nd prize in ImageNet'13
4. 1st prize in ImageNet'14

Best paper/demo awards:
ACM MM13,
ACM MM12,
Also licensed to ******

LFW: 99.70%, among best two
Best human parsing performance
Cross-age synthesis
Face analysis with occlusions

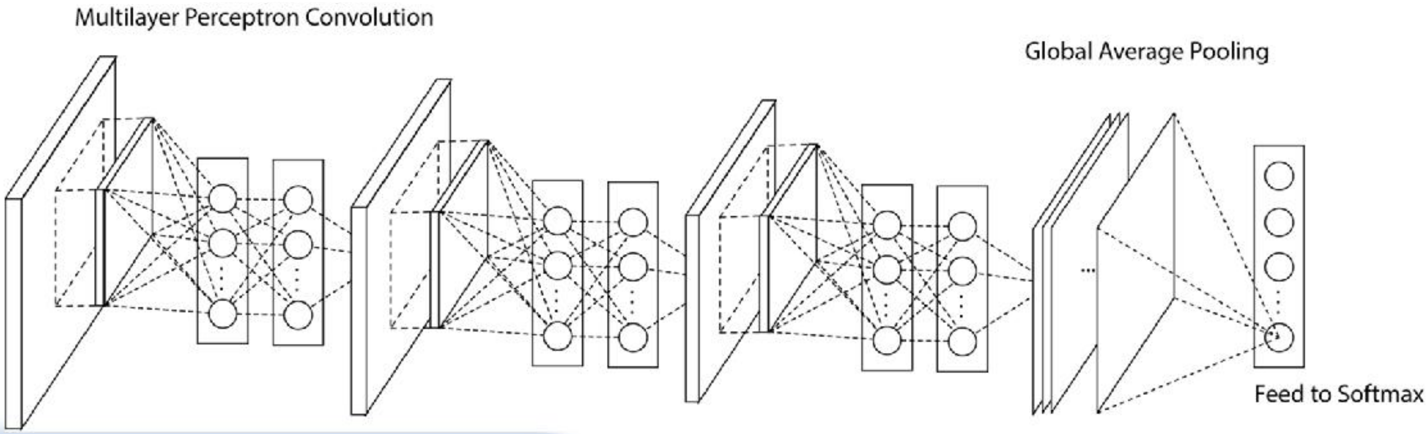
Big Data Analytics:
Intelligent Recommendation
Inventory Planning
Assistive driving

Task I: Face Recognition

(Network-in-Network)

"Network in Network" (NIN)

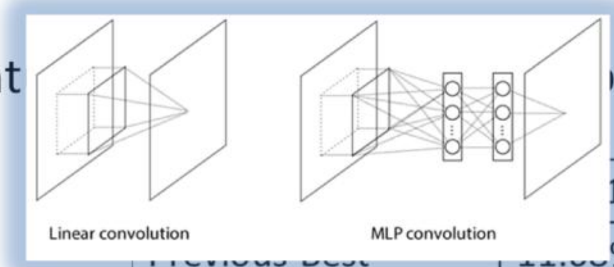
NIN: more brain-like || complex-cell filters, pure convolutional



NIN

feature maps = # classes

- Int ally, and more discriminative locally



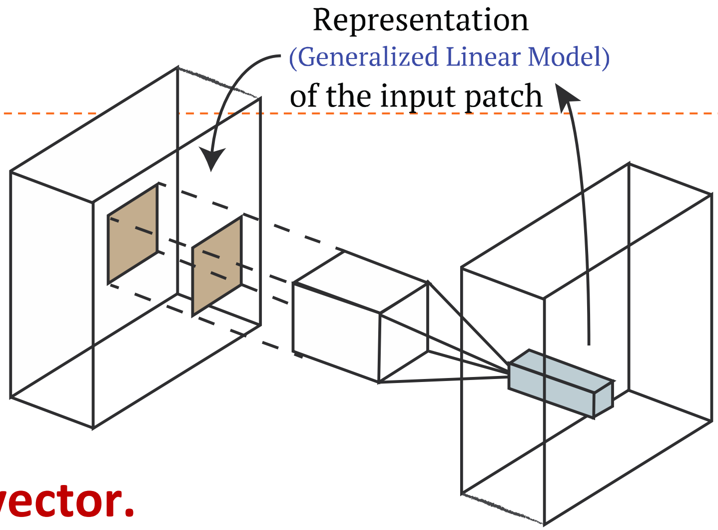
	10	Cifar-100
Previous Best	%	38.57%
	%	36.30%

Can be any small networks, e.g. MLP, Inception module, batch-normalization, or others for other particular targets, but **SMALL**

With less parameter #

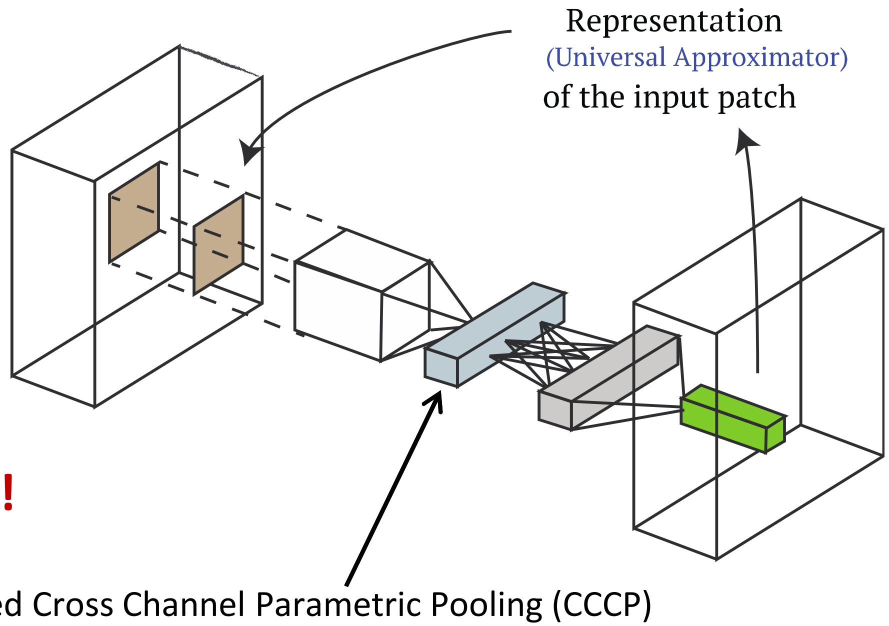
[4] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C. Courville, Yoshua Bengio: Maxout Networks. ICML (3) 2013: 1319-1327

Better Local Abstraction



$$y = \phi(w^T x + b)$$

Local patch is projected to its feature vector.
Using a small network.



$$y_i = \phi(w_i^T y_{i-1} + b_i)$$

$$y_0 = x$$

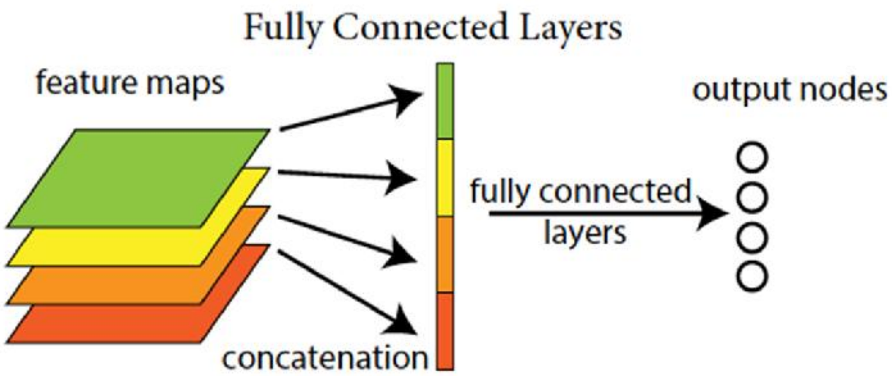
Motivation: Better Local Abstraction!

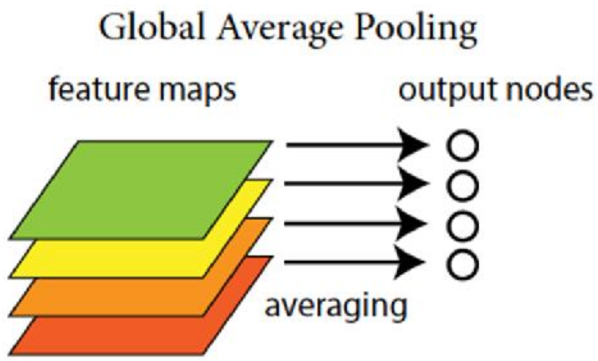
Lin, Min, Qiang Chen, and Shuicheng Yan. "Network In Network." ICLR-2014.

Much Smaller Model

CNN

NIN

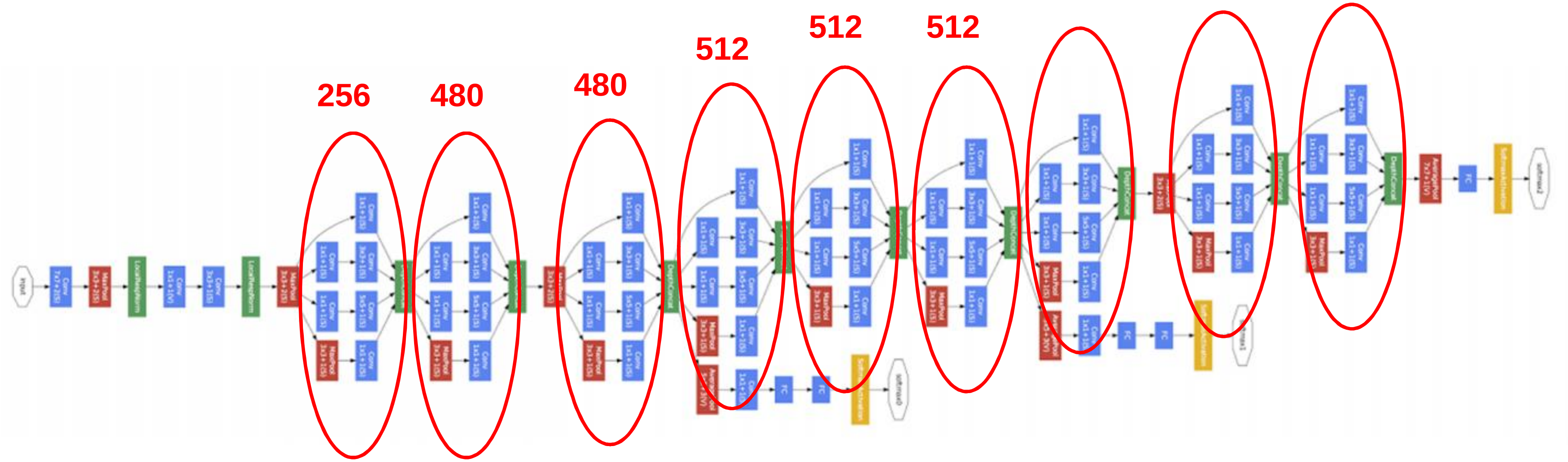




Confidence map of each category

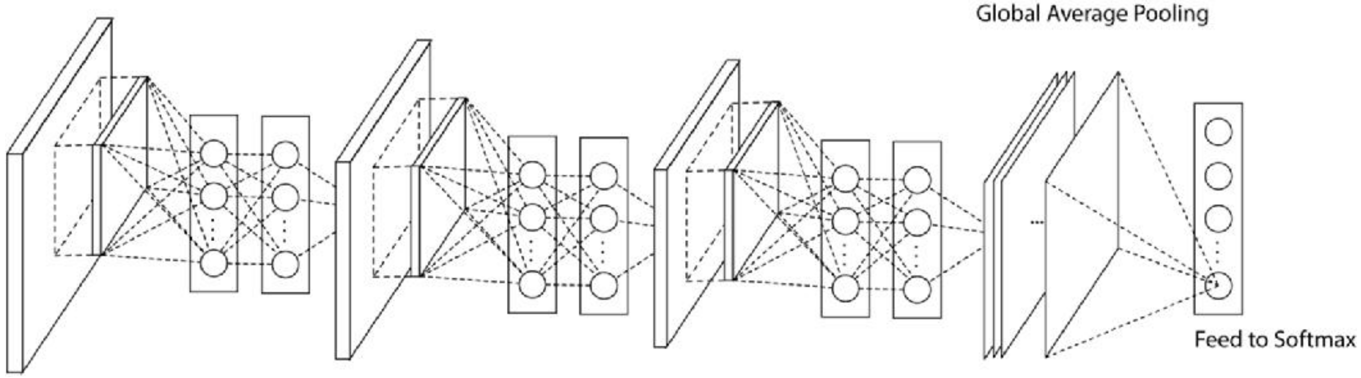
Save tons of parameters

Inspiring other Deeper Models



GoogLeNet = Deeper Network-in-Network

Face Recognition with Deeper NINs



- Deeper NINs trained over 494k images of 10k subjects [from Prof. Stan Li's group], followed by binary classifier
- Current accuracy on LFW is **99.70%** **(said to the reasonable upper-bound)**

Organization	Accuracy
Baidu	99.62%-> 99.82%
Face++	99.50%
CUHK	99.47%
Facebook	98.37%

LFW

NUS National University of Singapore

A Face Recognition Story

WeChat

Her son said they are Mummy and Daddy!
Cross-border officers often challenged her!

? Same Person ?

Our system answers “**Yes**”, Distance = 8 < Threshold = 200

Task 2: Human Parsing

(Fully-convolutional Network)

Task: Human Parsing

- Decompose a human photo into semantic fashion/body items
- Pixel-level semantic labeling

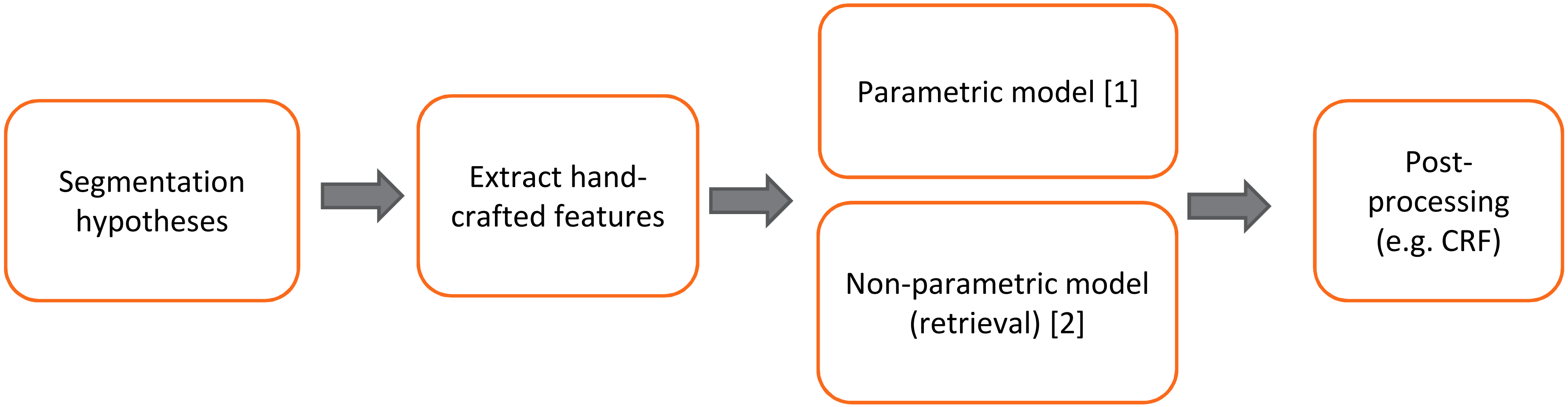
Human Parsing = Engine for Applications

Person Re-ID

State-of-the-art Related Solutions

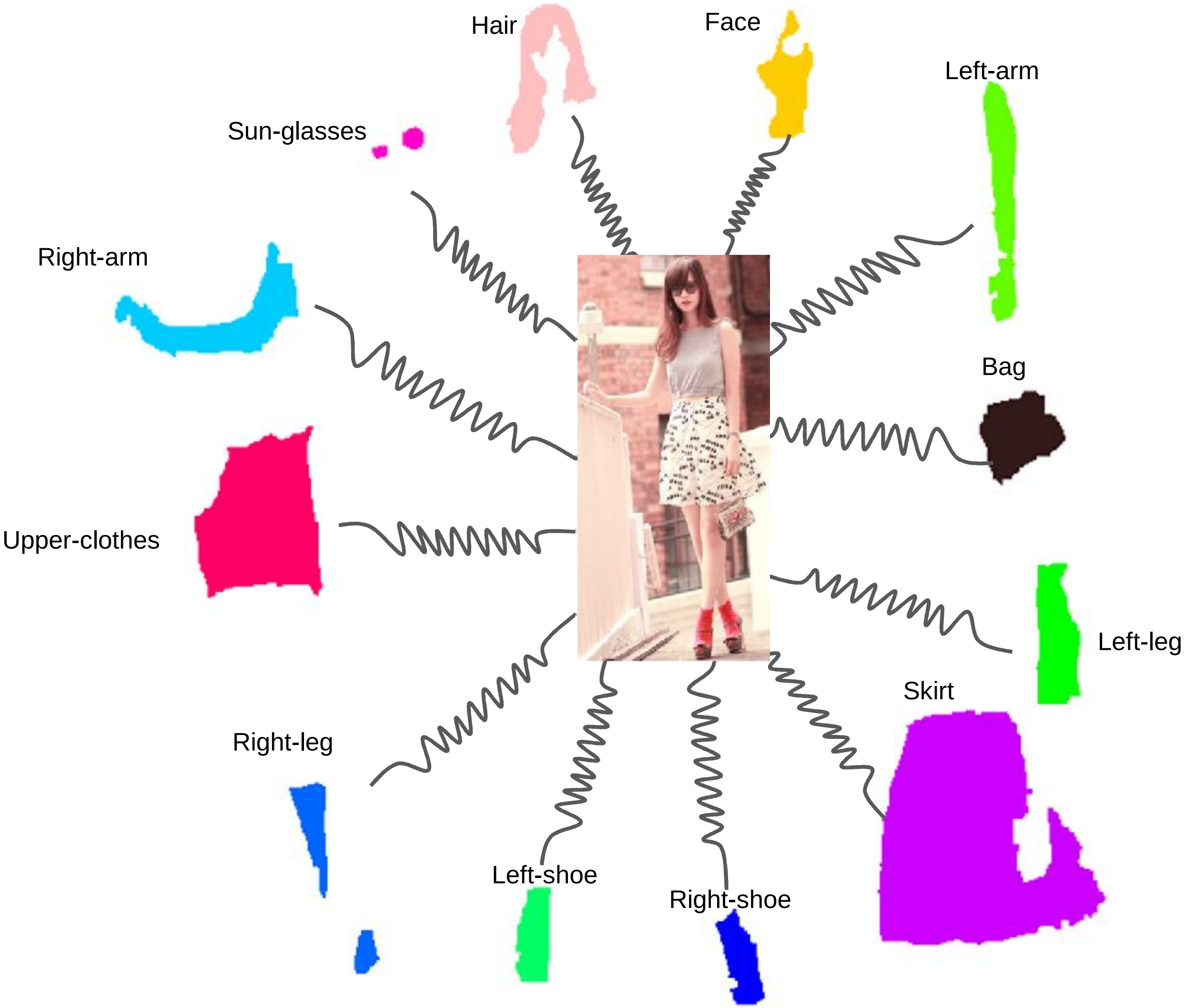
- ## Hand-designed pipelines

--- Heavily rely on the performance of individual component
--- Founded on hand-designed features and complex context models



[1] Jian Dong, Qiang Chen, Wei Xia, ZhongYang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. In ICCV, 2013
[2] K. Yamaguchi, M.H. Kiapour, and T.L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In ICCV, 2013

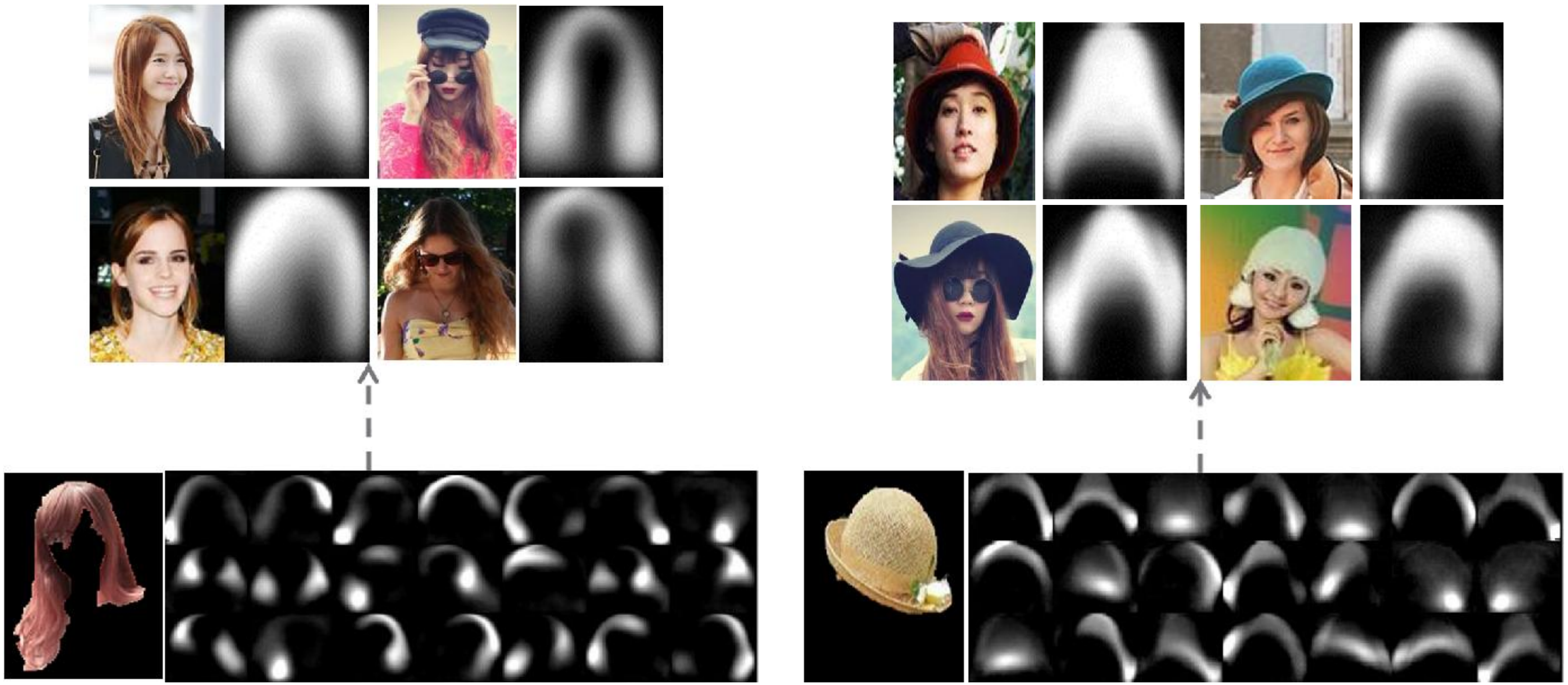
Motivation



- Deformable Human Items Model (similar to ASM): predict the **normalized item masks**, and their **active shape/location parameters** with two CNN networks

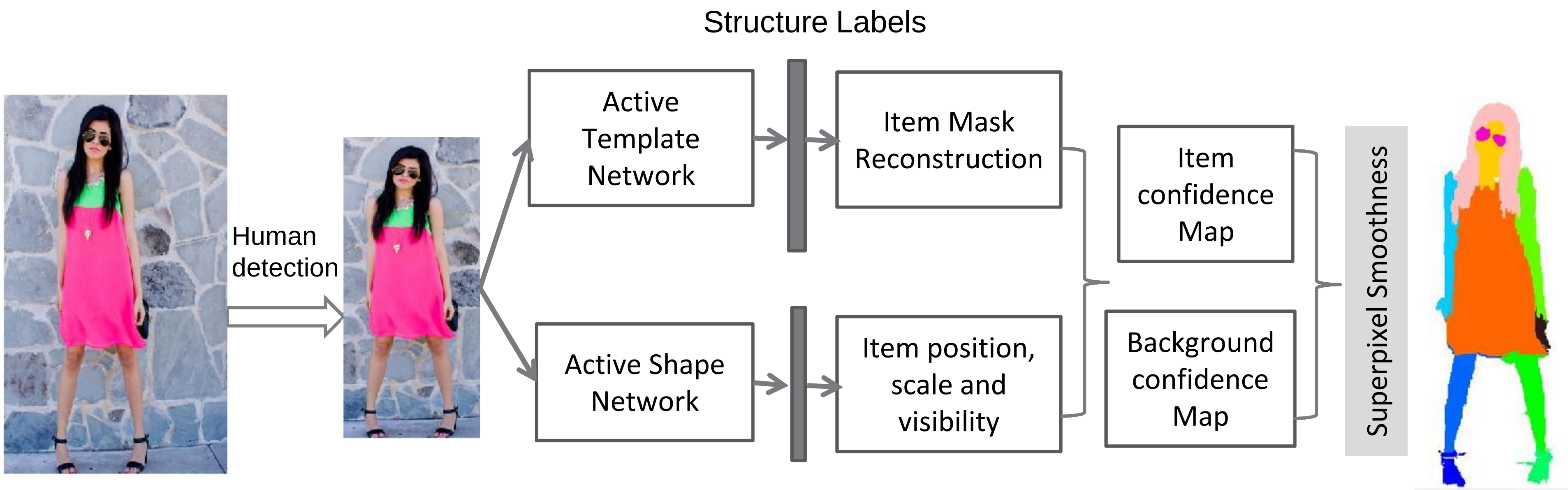
Normalized Item Mask

- The masks of different items often appear in various specific shapes
- The mask can be approximated as a linear combination of the learned templates



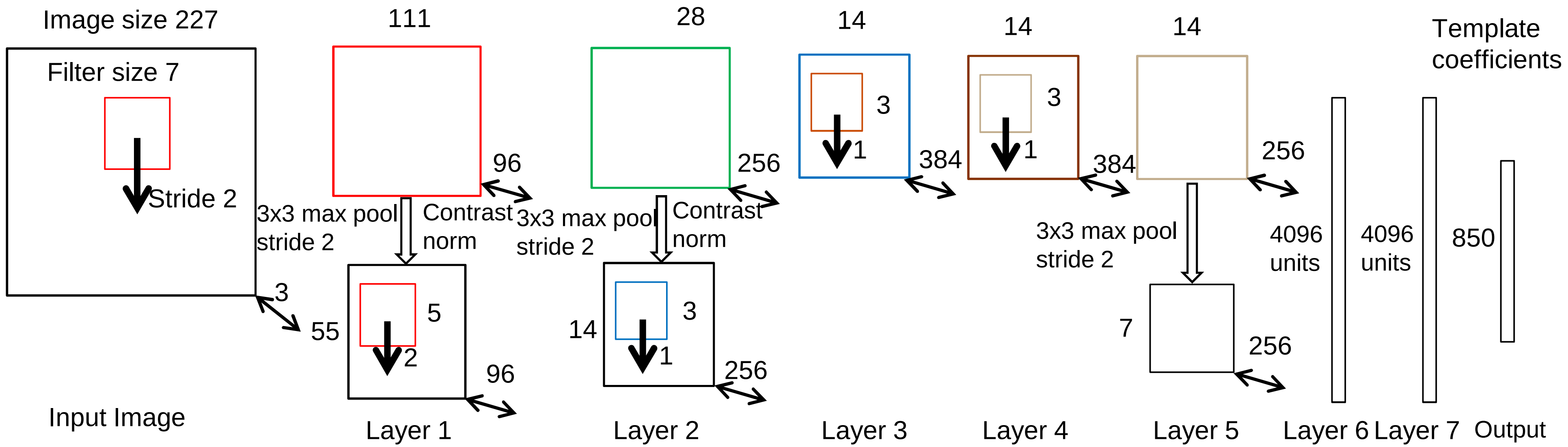
Our Framework

- Active Template Network for predicting item template coefficients
- Active Shape Network for predicting active shape/location parameters
- Combine the resulting structure outputs and then refine the parsing result



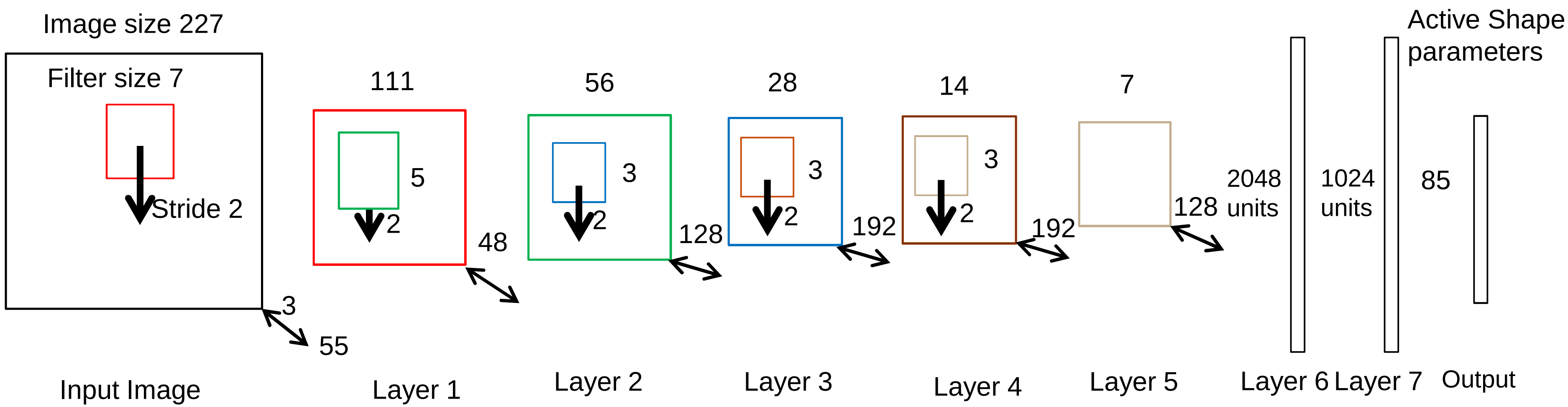
Active Template Network

- Learn 50 templates for each item by Non-negative Matrix Factorization (NMF) in an offline way
- Regress the output: 50*17 for 17 human items



Active Shape Network

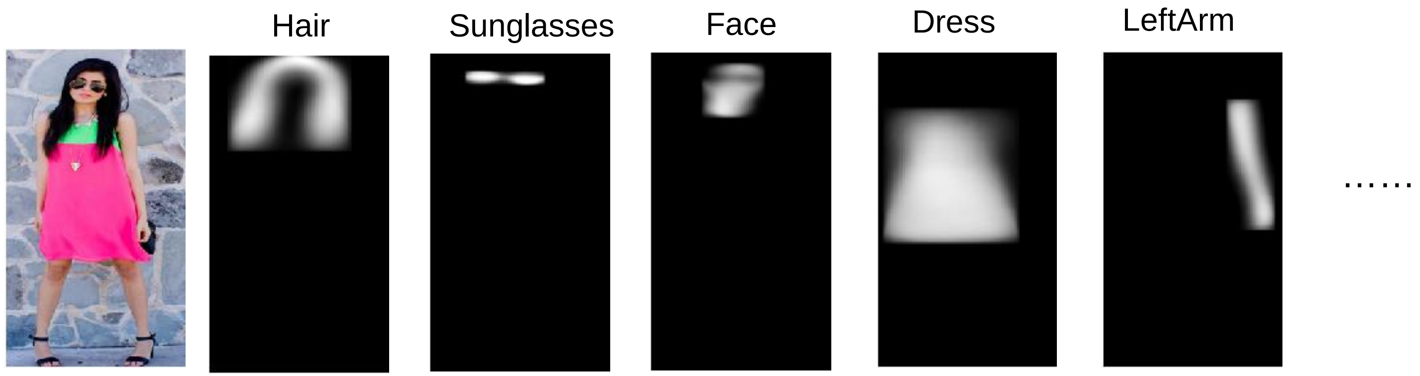
- Predict x,y coordinates, width, height, visibility flag for each item
- Eliminate the max-pooling layer in CNN to keep the position sensitiveness



	Accuracy	Foreground accuracy	Average precision	Average recall	Average F-1 scores
Original Structure	90.21	67.17	69.16	56.04	60.77
Ours	91.01	70.40	69.61	58.82	62.78

Structure Output Combination

- Combine the structure outputs from two networks, and generate 17 confidence maps of the human items



- Optional bounding-box refinement and super-pixel smoothening

Results

- Datasets: 7,700 images, 6,000 for training, 1,000 for testing and 700 for validation

- Training: Manually decrease the learning rate according to the validation error

- Training time: for 120 epochs, take 2-3 days on two NVIDIA GTX TITAN 6GB GPUs

- Testing time: process one image within about 0.5 second

NUS National University of Singapore

Results

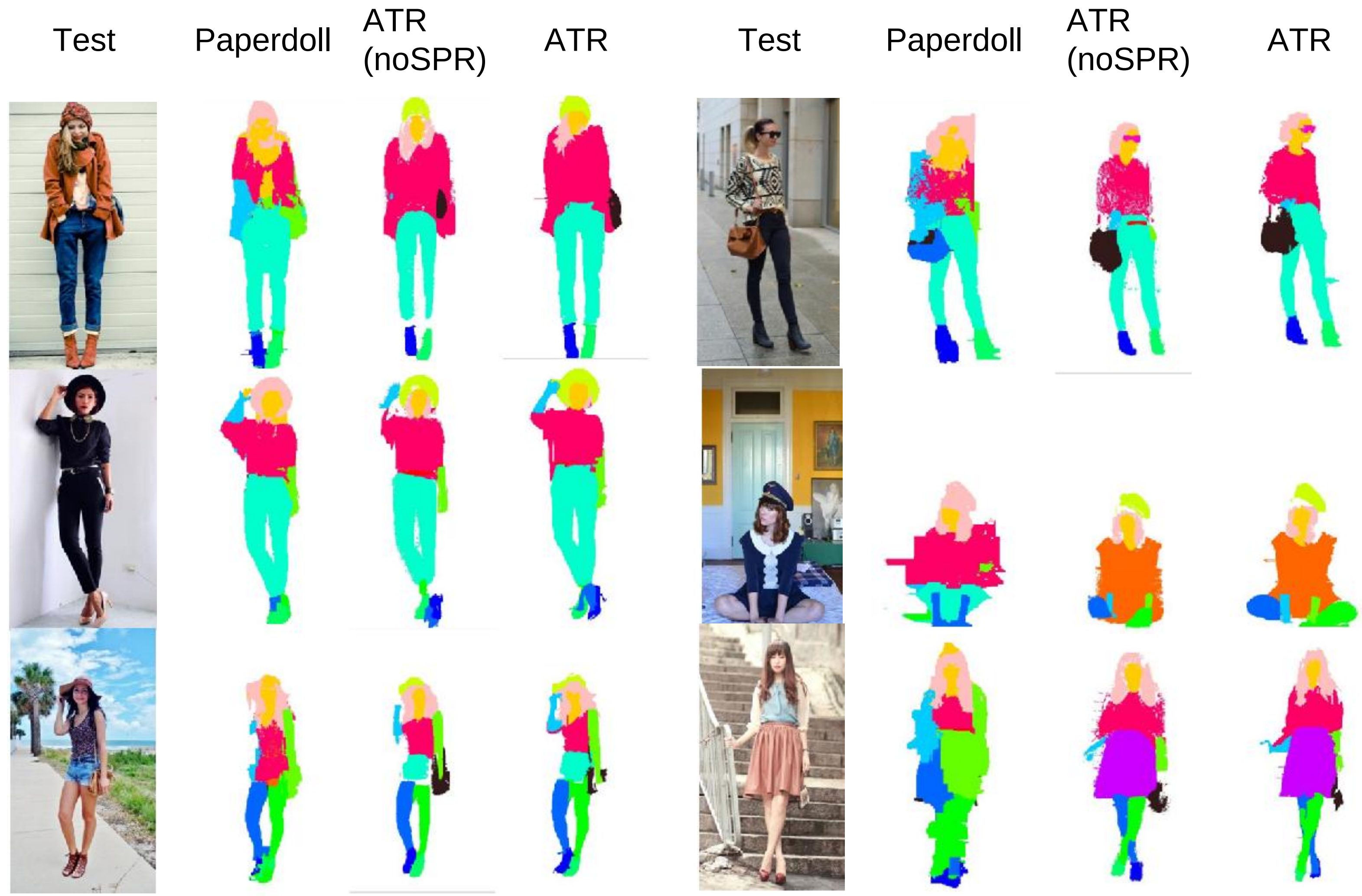
Comparison of parsing performances with two state-of-the-art methods:

	Accuracy	Foreground accuracy	Average precision	Average recall	Average F-1 scores
Yamaguchi [3]	84.38	55.59	37.54	51.05	41.80
Paper-doll [2]	88.96	62.18	52.75	49.43	44.76
ATR(noSPR)	89.33	64.79	63.75	56.19	59.60
ATR	91.01	70.40	69.16	58.82	62.78
ATR + BBox Regression	**91.11**	**71.04**	**71.69**	**60.25**	**64.38**

[2] K. Yamaguchi, M.H. Kiapour, and T.L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In ICCV, 2013

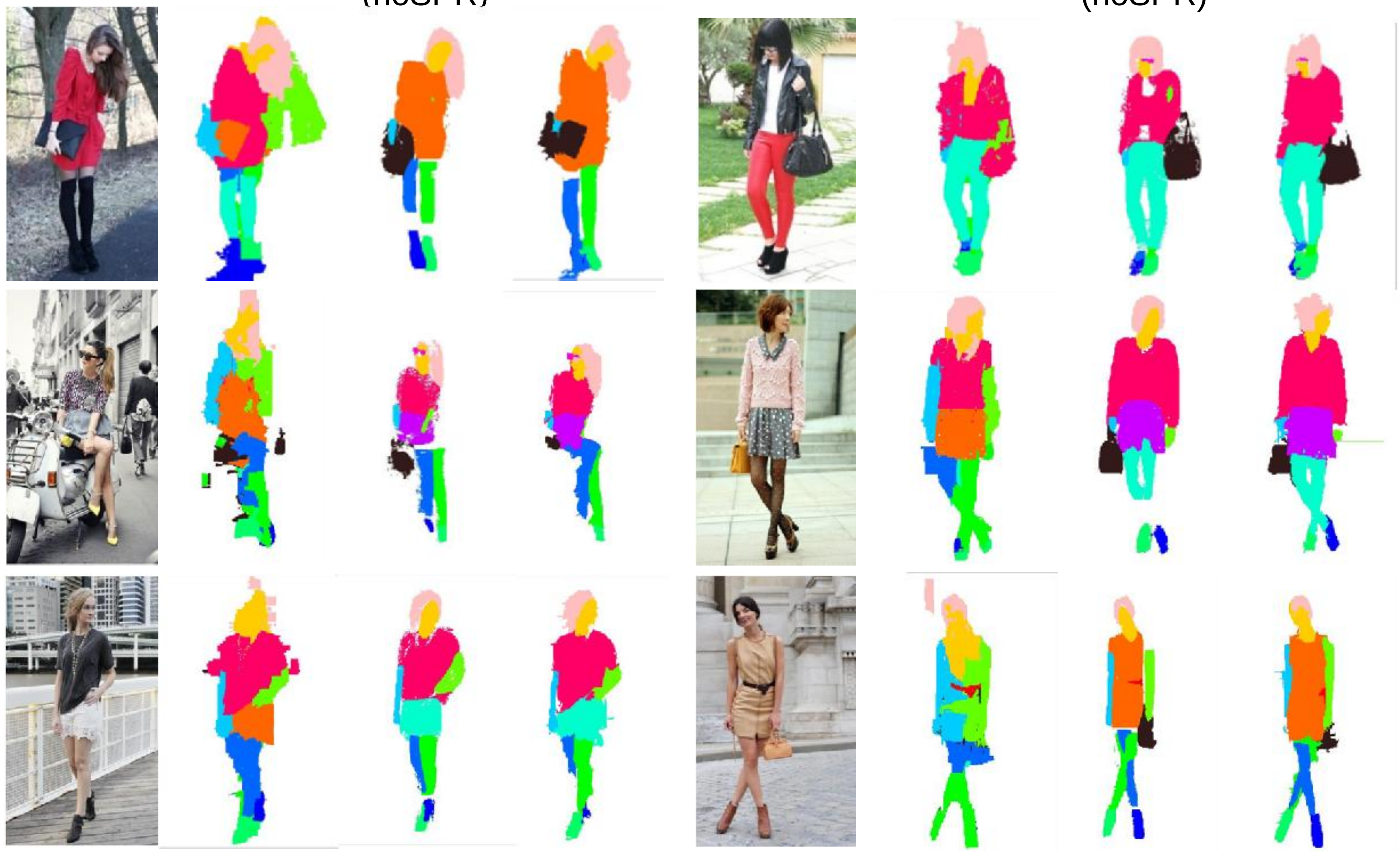
[3] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, and T.L. Berg. Parsing clothing in fashion photographs. In CVPR 2012.

Parsing Results



Parsing Results

Test | Paperdoll | ATR (noSPR) | ATR | Test | Paperdoll | ATR (noSPR) | ATR



Limitations

- Two separate networks lead to sub-optimal results
- Results are still with many artifacts
- Super-pixel smoothing are performed as post-processing step

Our new framework improves the mAP from 64.38% to 76.95%:

Human Parsing with Contextualized Convolutional Neural Network

Motivations

- Integrate multi-source contexts into a fully convolutional network

✓ Cross-layer context:

: multi-level feature fusion

✓ Global image-level context:

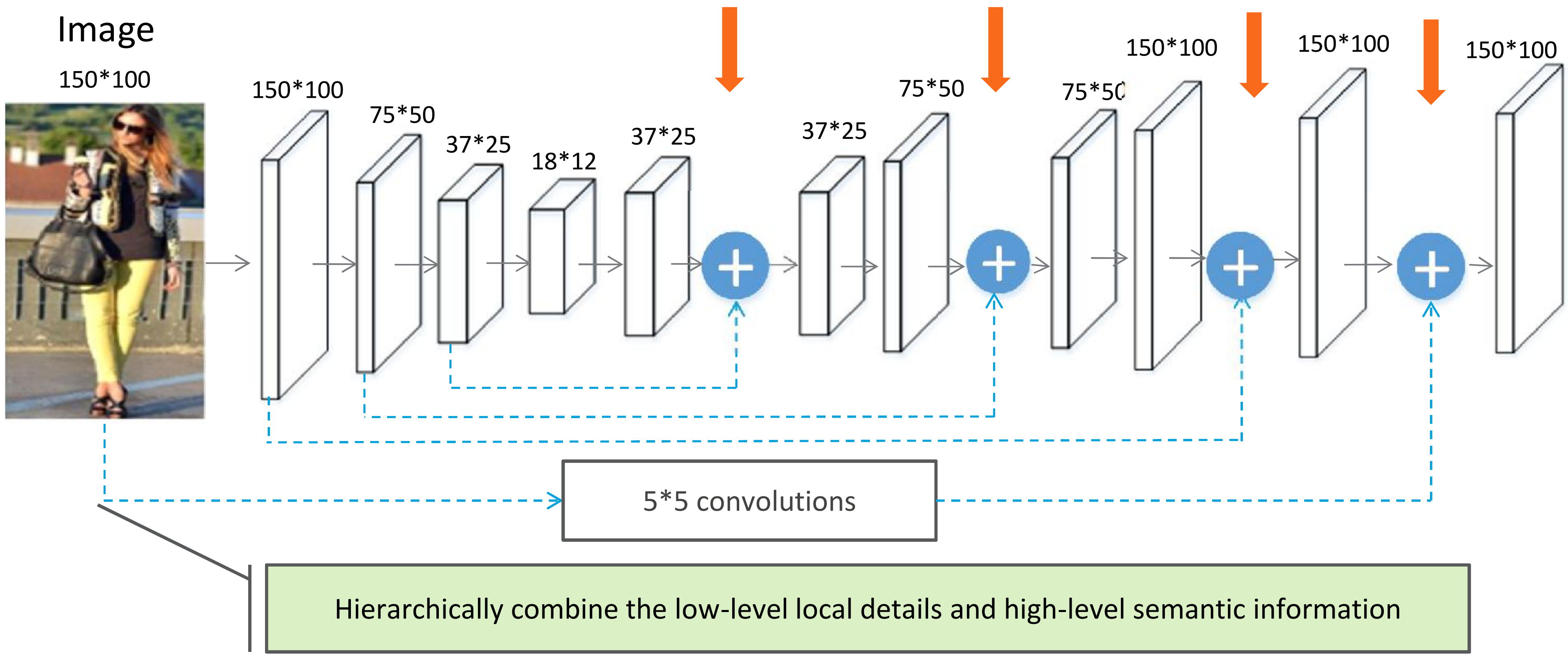
: coherence between pixel-wise labelling and image label prediction

✓ Local Super-pixel context:

: local boundaries and label consistency among similar neighbouring super-pixels

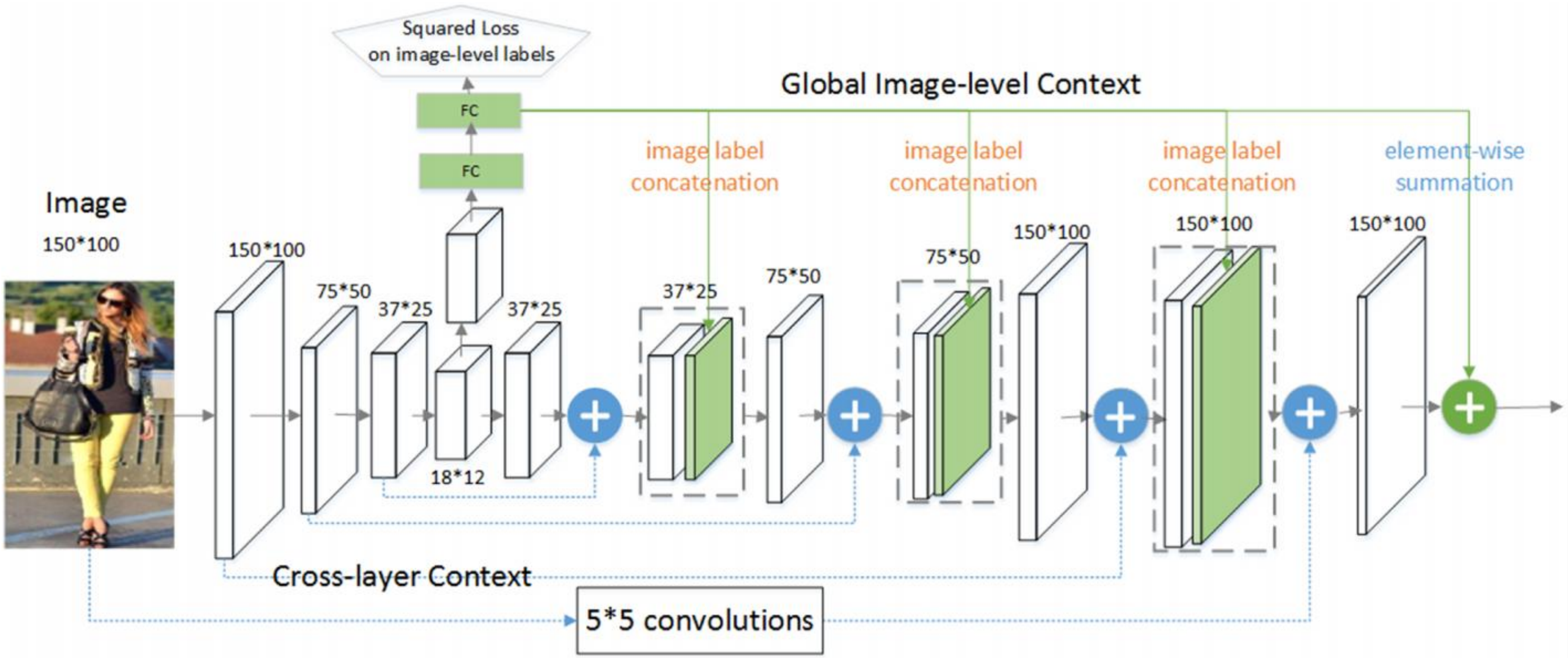
Contextualized Network

- Cross-layer context
  - Four feature map fusions



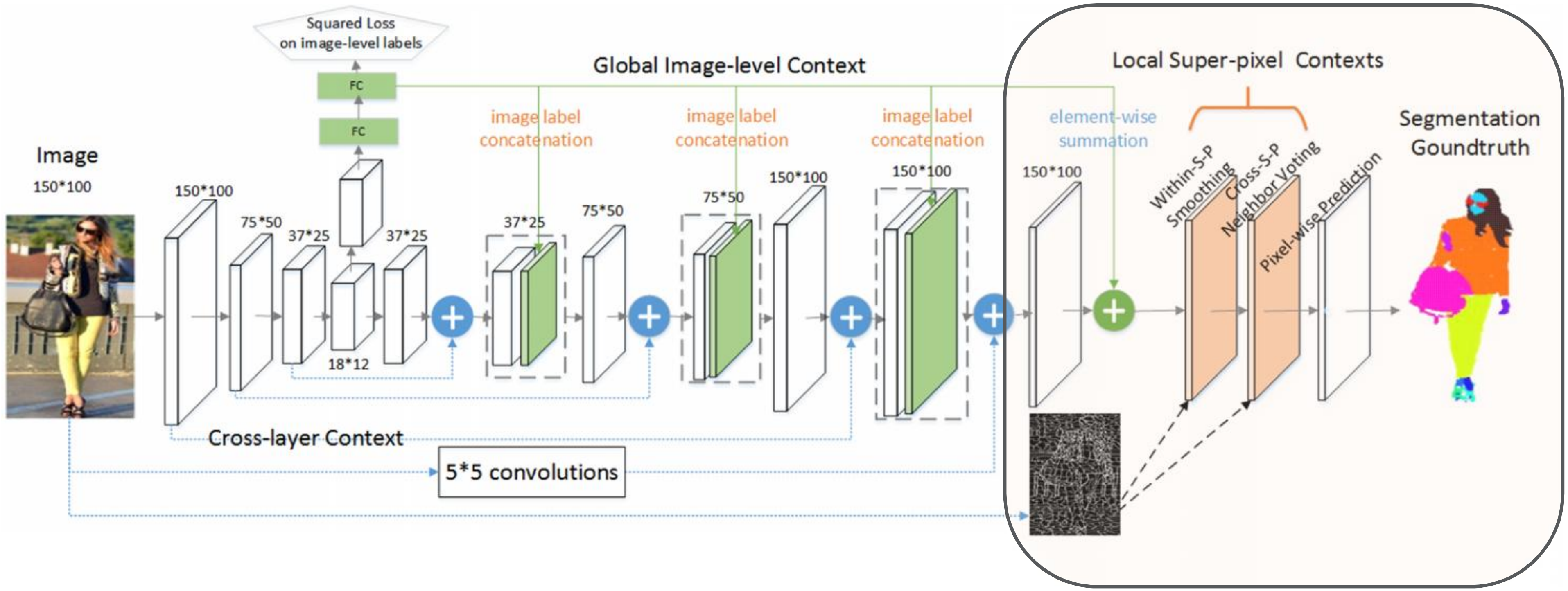
Contextualized Network

- Global image-level context
  - Incorporate global image label prediction



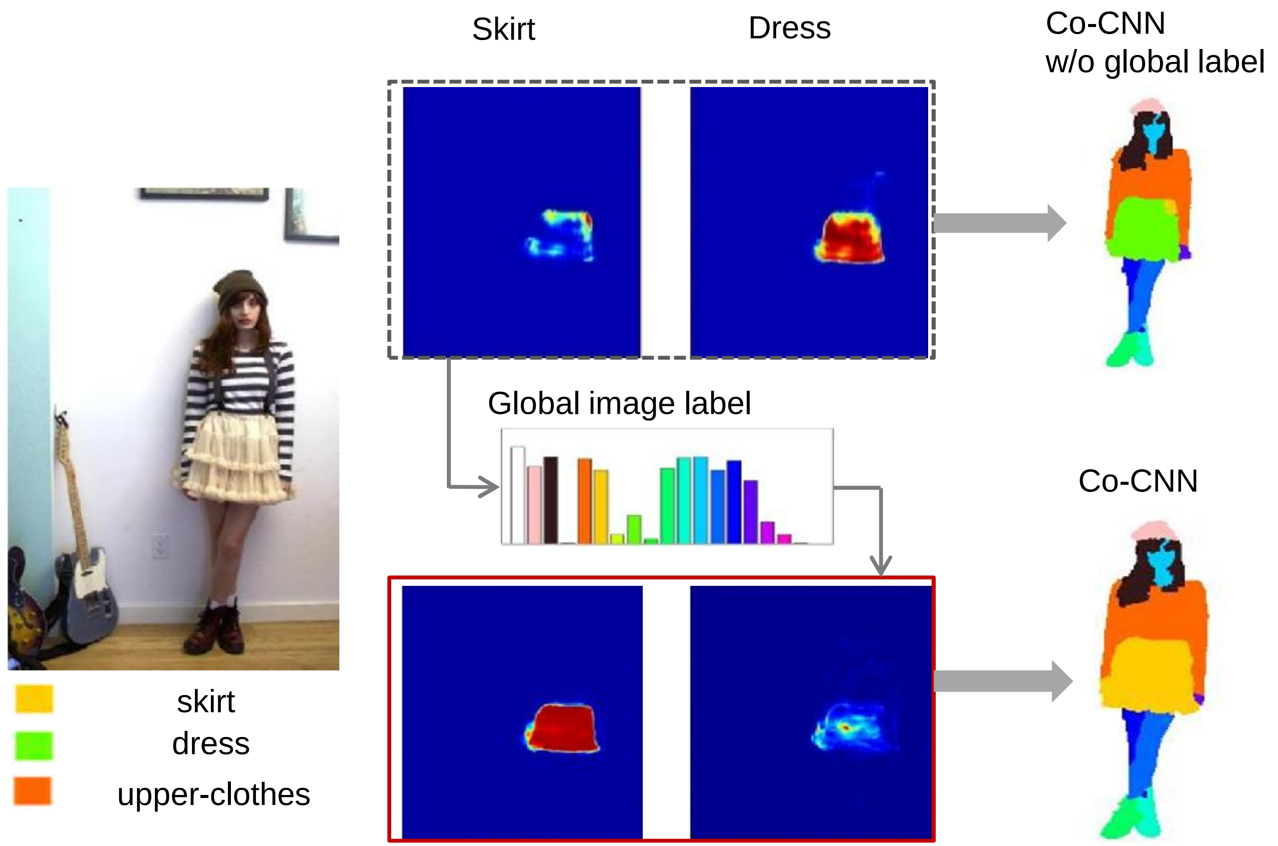
Contextualized Network

- Local Super-pixel context
  - Integrate within-super-pixel smoothing and cross-super-pixel neighbourhood voting



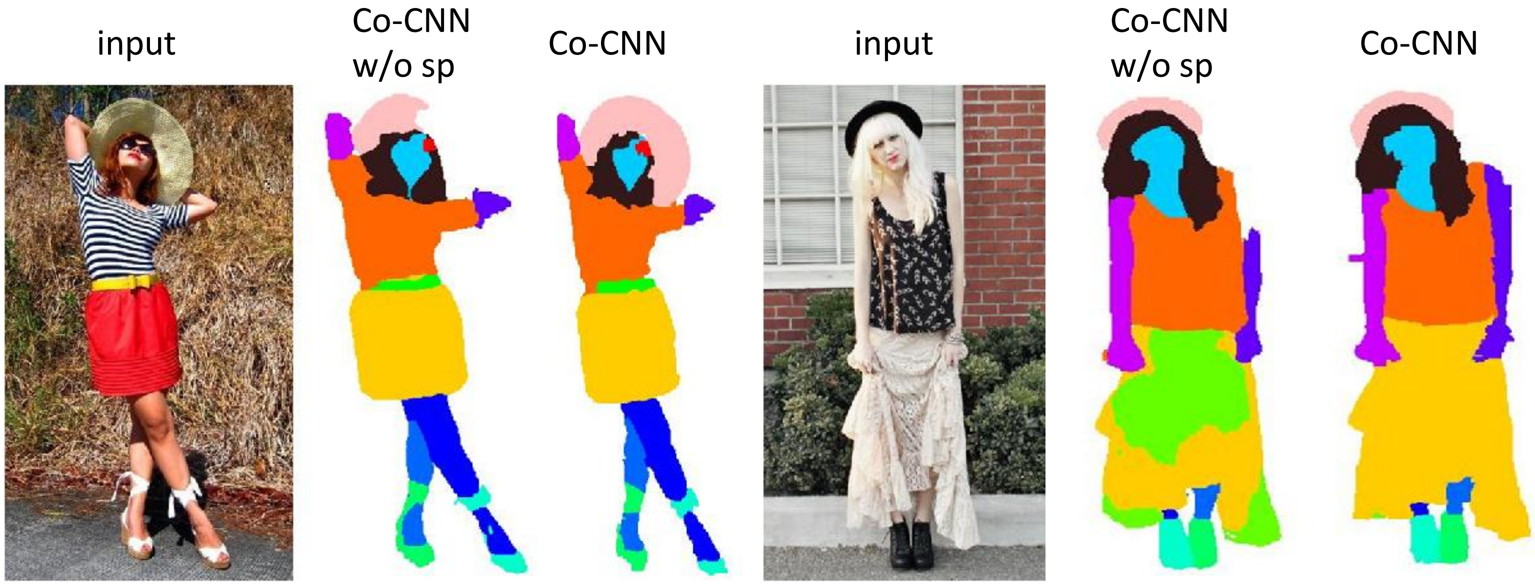
Contextualized Network

- Global image-level context helps distinguish the ambiguous labels



Contextualized Network

- Local super-pixel context retains the local boundaries and appearance consistency



Results

Comparison of parsing performances with four state-of-the-art methods on ATR dataset:

	Accuracy	Foreground accuracy	Average precision	Average recall	Average F-1 scores
Yamaguchi et ak.	84.38	55.59	37.54	51.05	41.80
Paperdoll	88.96	62.18	52.75	49.43	44.76
M-CNN	89.57	73.98	64.56	65.17	62.81
ATR	91.11	71.04	71.69	60.25	64.38
Co-CNN	**95.23**	**80.90**	**81.55**	**74.42**	**76.95**

Results

- Analyses on architectural variants of our model

Cross-layer context

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [28]	84.38	55.59	37.54	51.05	41.80
⋆ PaperDoll [27]	88.96	62.18	52.75	49.43	44.76
⋆ M-CNN [18]	89.57	73.98	64.56	65.17	62.81
⋆ ATR [15]	91.11	71.04	71.69	60.25	64.38
baseline (150-75)	92.77	68.66	67.98	62.85	63.88
baseline (150-75-37)	92.91	76.29	78.48	65.42	69.32
baseline (150-75-37-18)	94.41	78.54	76.62	71.24	72.72
⋆ baseline (150-75-37-18, post-process)	94.48	78.85	77.22	71.78	73.25
baseline (150-75-37-18, w/o fusion)	92.57	70.76	67.17	64.34	65.25
baseline (150-75-37-18, lessfilters)	94.23	77.79	75.66	70.42	71.82
baseline (150-75-37-18, concat)	93.10	72.17	69.63	66.94	67.82
Co-CNN (concatenate with global label)	94.90	80.80	78.35	73.14	74.56
Co-CNN (concatenate, summation with global label)	94.87	79.86	78.00	73.94	75.27
Co-CNN (w-s-p)	95.09	80.50	79.22	74.38	76.17
Co-CNN (full)	95.23	80.90	81.55	74.42	76.95

Results

- Analyses on architectural variants of our model

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [28]	84.38	55.59	37.54	51.05	41.80
⋆ PaperDoll [27]	88.96	62.18	52.75	49.43	44.76
⋆M-CNN [18]	89.57	73.98	64.56	65.17	62.81
⋆ ATR [15]	91.11	71.04	71.69	60.25	64.38
baseline (150-75)	92.77	68.66	67.98	62.85	63.88
baseline (150-75-37)	92.91	76.29	78.48	65.42	69.32
baseline (150-75-37-18)	94.41	78.54	76.62	71.24	72.72
(150-75-37-18, post-process)	94.48	78.85	77.22	71.78	73.25
(150-75-37-18, w/o fusion)	92.57	70.76	67.17	64.34	65.25
(150-75-37-18, lessfilters)	94.23	77.79	75.66	70.42	71.82
baseline (150-75-37-18, concat)	93.10	72.17	69.63	66.94	67.82
Co-CNN (concatenate with global label)	94.90	80.80	78.35	73.14	74.56
Co-CNN (concatenate, summation with global label)	94.87	79.86	78.00	73.94	75.27
Co-CNN (w-s-p)	95.09	80.50	79.22	74.38	76.17
Co-CNN (full)	95.23	80.90	81.55	74.42	76.95

Global image label context

Results

- Analyses on architectural variants of our model

Method	Accuracy	F.g. accuracy	Avg. precision	Avg. recall	Avg. F-1 score
⋆ Yamaguchi et al. [28]	84.38	55.59	37.54	51.05	41.80
⋆ PaperDoll [27]	88.96	62.18	52.75	49.43	44.76
⋆M-CNN [18]	89.57	73.98	64.56	65.17	62.81
⋆ ATR [15]	91.11	71.04	71.69	60.25	64.38
baseline (150-75)	92.77	68.66	67.98	62.85	63.88
baseline (150-75-37)	92.91	76.29	78.48	65.42	69.32
baseline (150-75-37-18)	94.41	78.54	76.62	71.24	72.72
⋆ baseline (150-75-37-18, post-process)	94.48	78.85	77.22	71.78	73.25
baseline (150-75-37-18, w/o fusion)	92.57	70.76	67.17	64.34	65.25
baseline (150-75-37-18, lessfilters)	94.23	77.79	75.66	70.42	71.82
50-75-37-18, concat)	93.10	72.17	69.63	66.94	67.82
tenate with global label)	94.90	80.80	78.35	73.14	74.56
e, summation with global label)	94.87	79.86	78.00	73.94	75.27
Co-CNN (w-s-p)	95.09	80.50	79.22	74.38	76.17
Co-CNN (full)	95.23	80.90	81.55	74.42	76.95

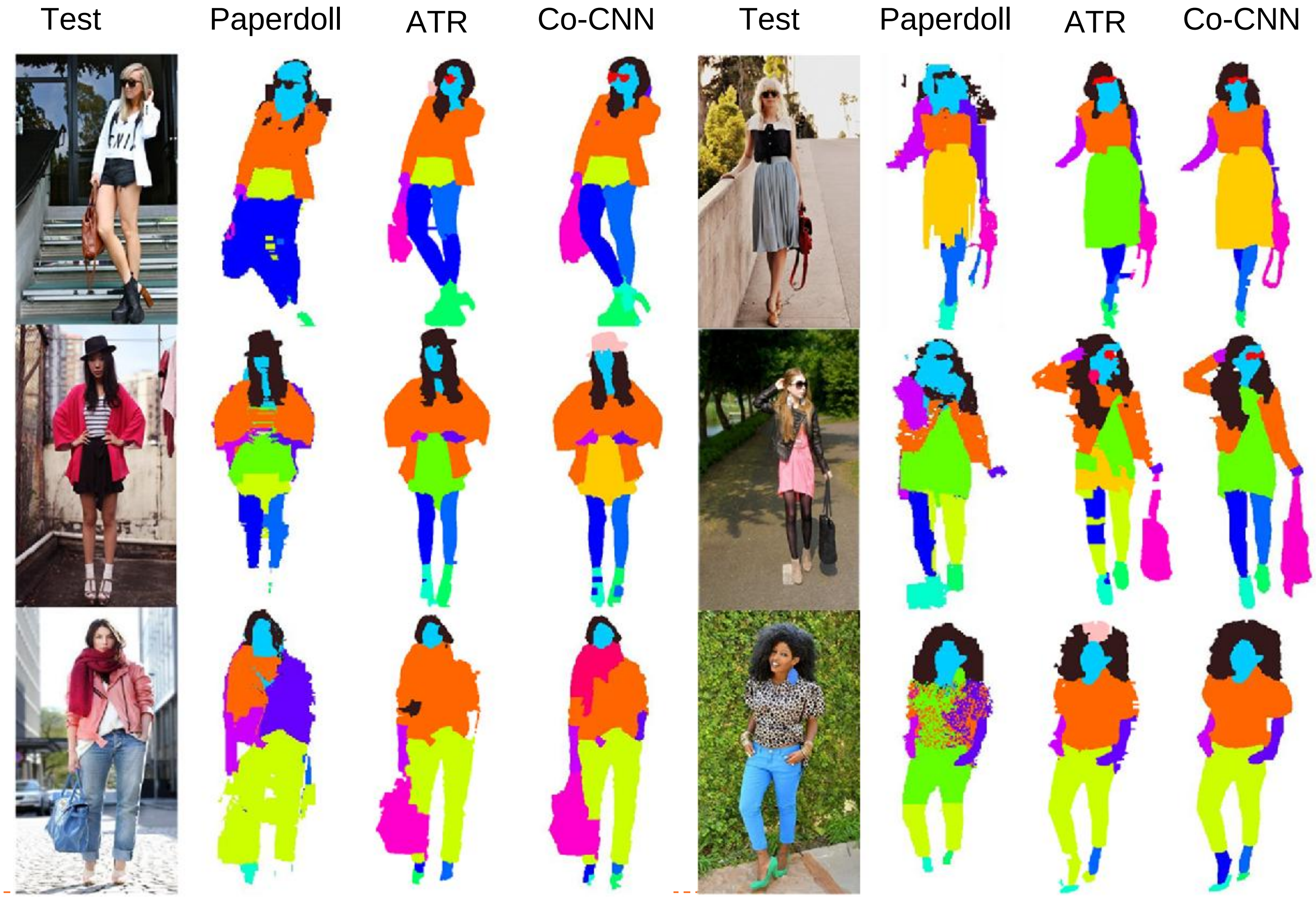
Local super-pixel context

Results

- Adding 10,000 human pictures from *"chictopia.com"*

	Accuracy	Foreground accuracy	Average precision	Average recall	Average F-1 scores
ATR	91.11	71.04	71.69	60.25	64.38
Co-CNN	95.23	80.90	81.55	74.42	76.95
Co-CNN(+Chictopia10k)	**96.02**	**83.57**	**84.95**	**77.66**	**80.14**

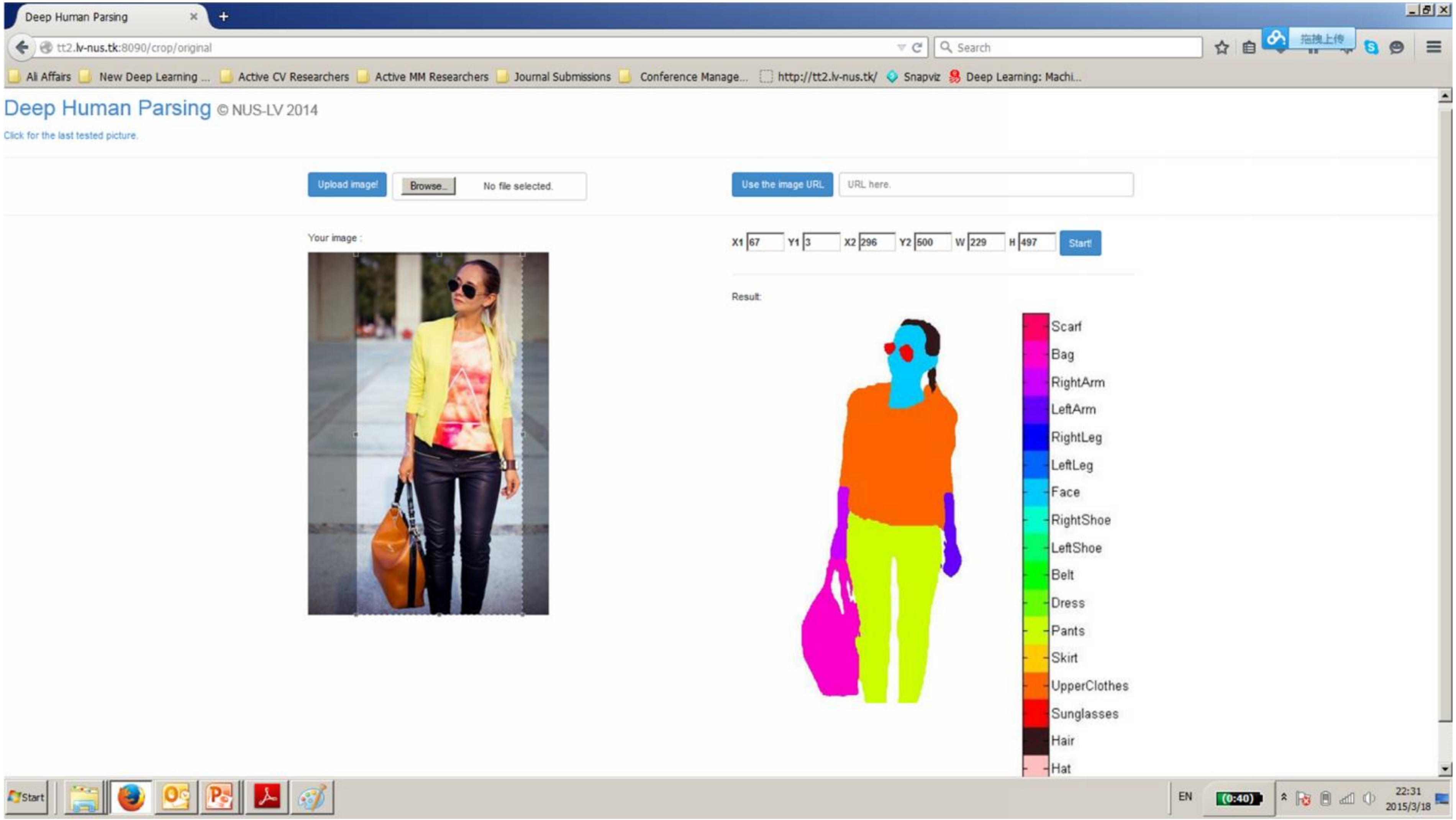
Parsing Results



Parsing Results

Test	Co-CNN w/o global label	Co-CNN	Test	Co-CNN w/o global label	Co-CNN

Online Human Parsing Engine (<0.15s)



Ready for many industry applications, including Re-ID ……

NUS National University of Singapore

Random Thoughts on Deep Learning for Biometrics

- Deep learning has shown great power for biometrics, so is the left issue “big data” or “new algorithm”?

- Industry is doing better than academia due to big data and computing resource, what should our academia focus?
  - Should we still focus on less-important research with small dataset or collaborate with industry?

- But anyway, good thing is that, more jobs and funding are there for us……

NUS
National University
of Singapore

Email:
eleyans@nus.edu.sg