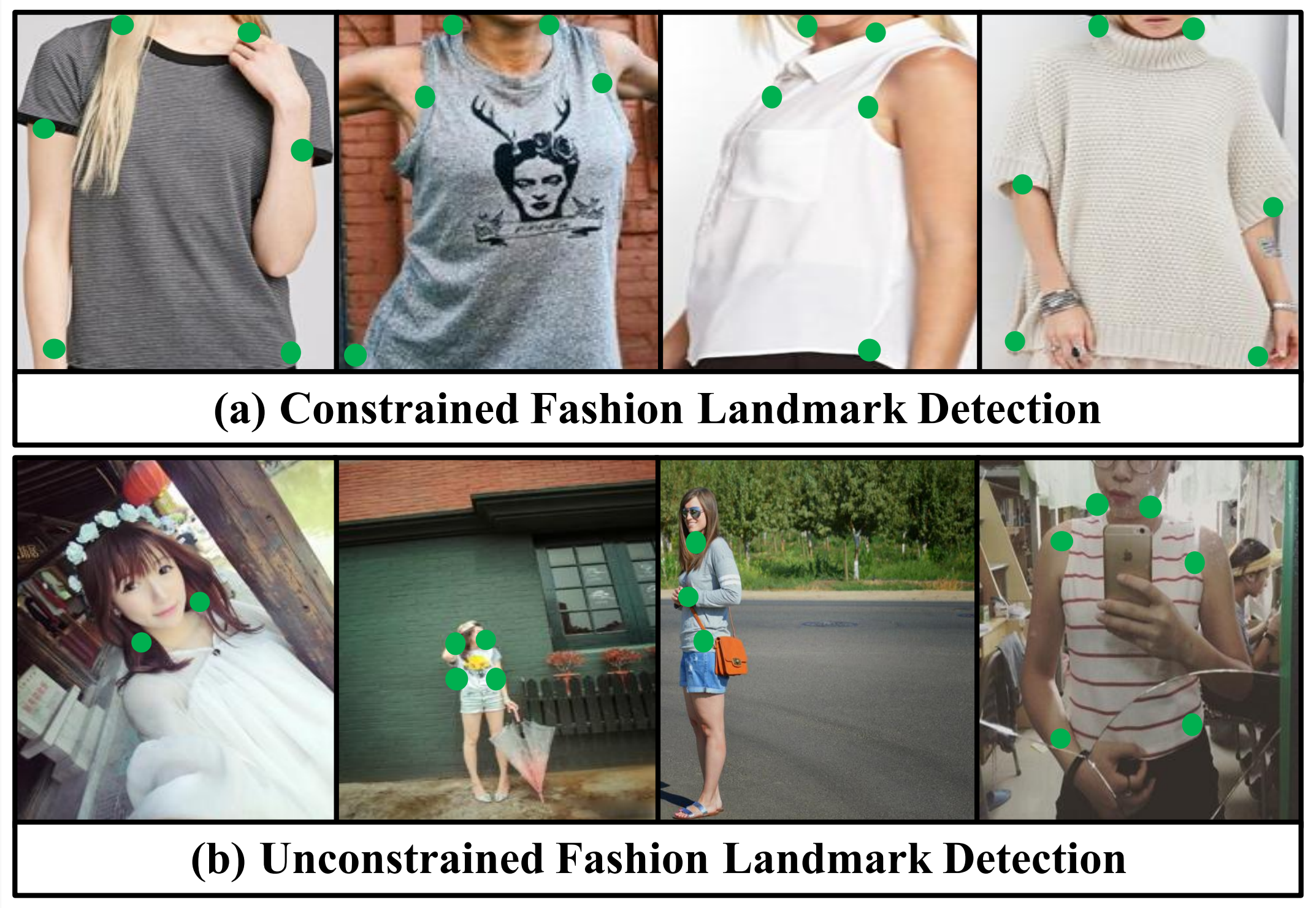
香港中文大學
The Chinese University of Hong Kong

Unconstrained Fashion Landmark Detection via Hierarchical Recurrent Transformer Networks

Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, Xiaoou Tang

ACMmultimedia25

Motivation


(a) Constrained Fashion Landmark Detection

(b) Unconstrained Fashion Landmark Detection

Problem

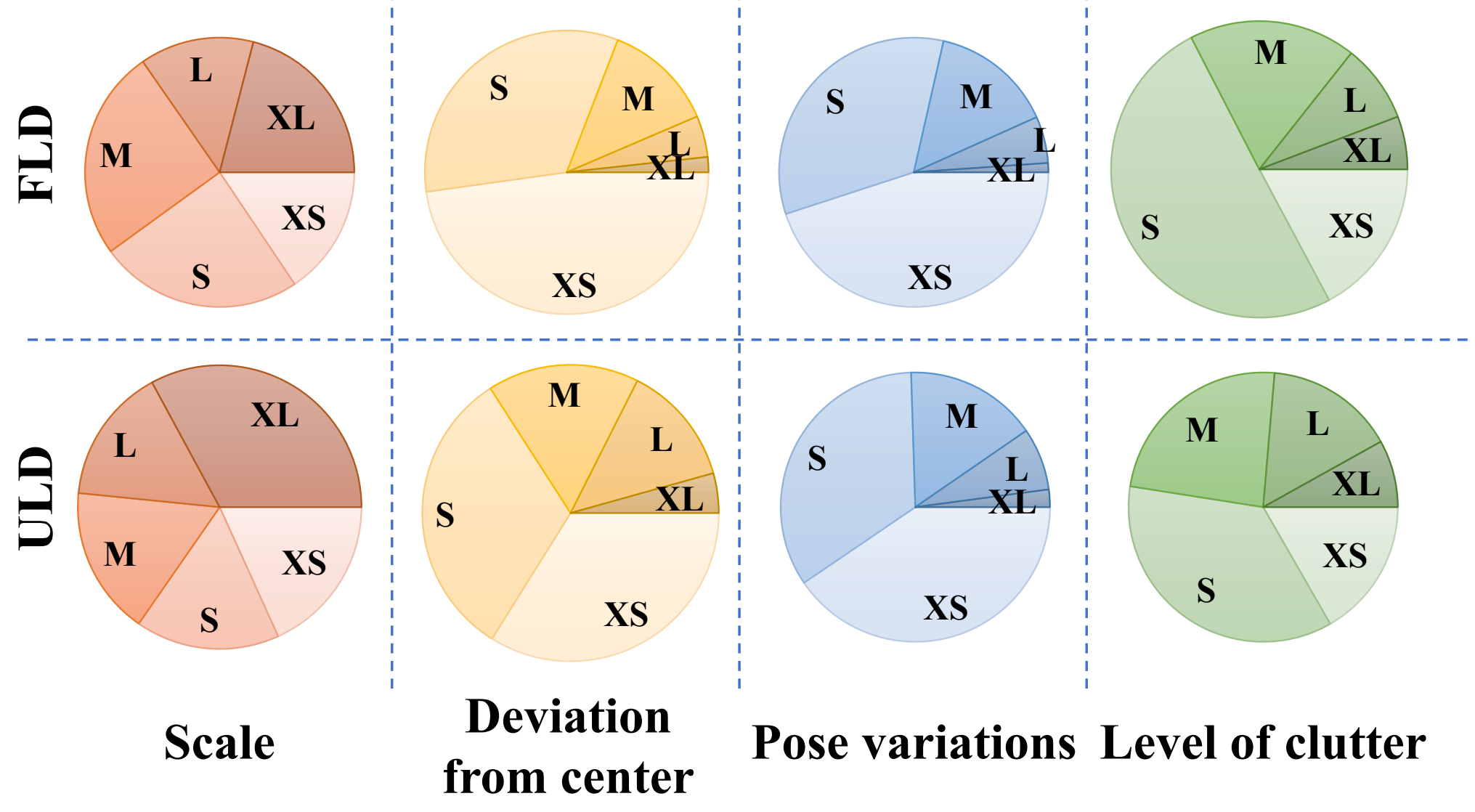
How to detect fashion landmarks without bounding boxes of clothes?

Difficulty

Background clutters, human poses, and scales variations due to:

- deformable objects,
- complex application scenarios.

Dataset

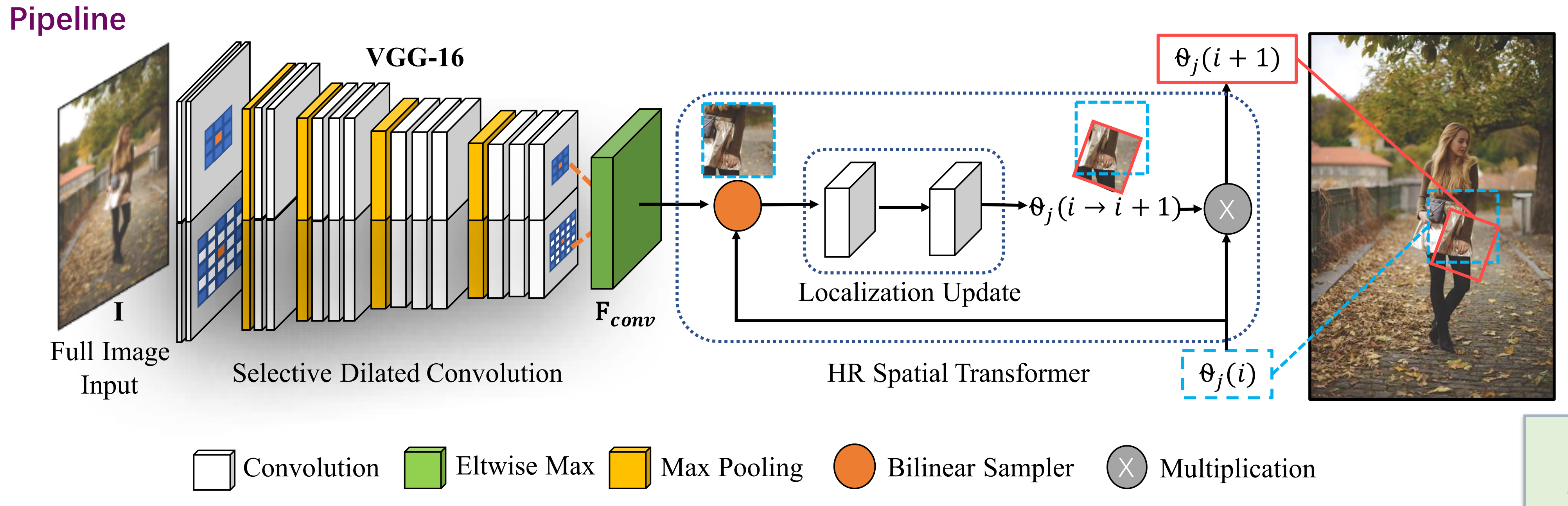


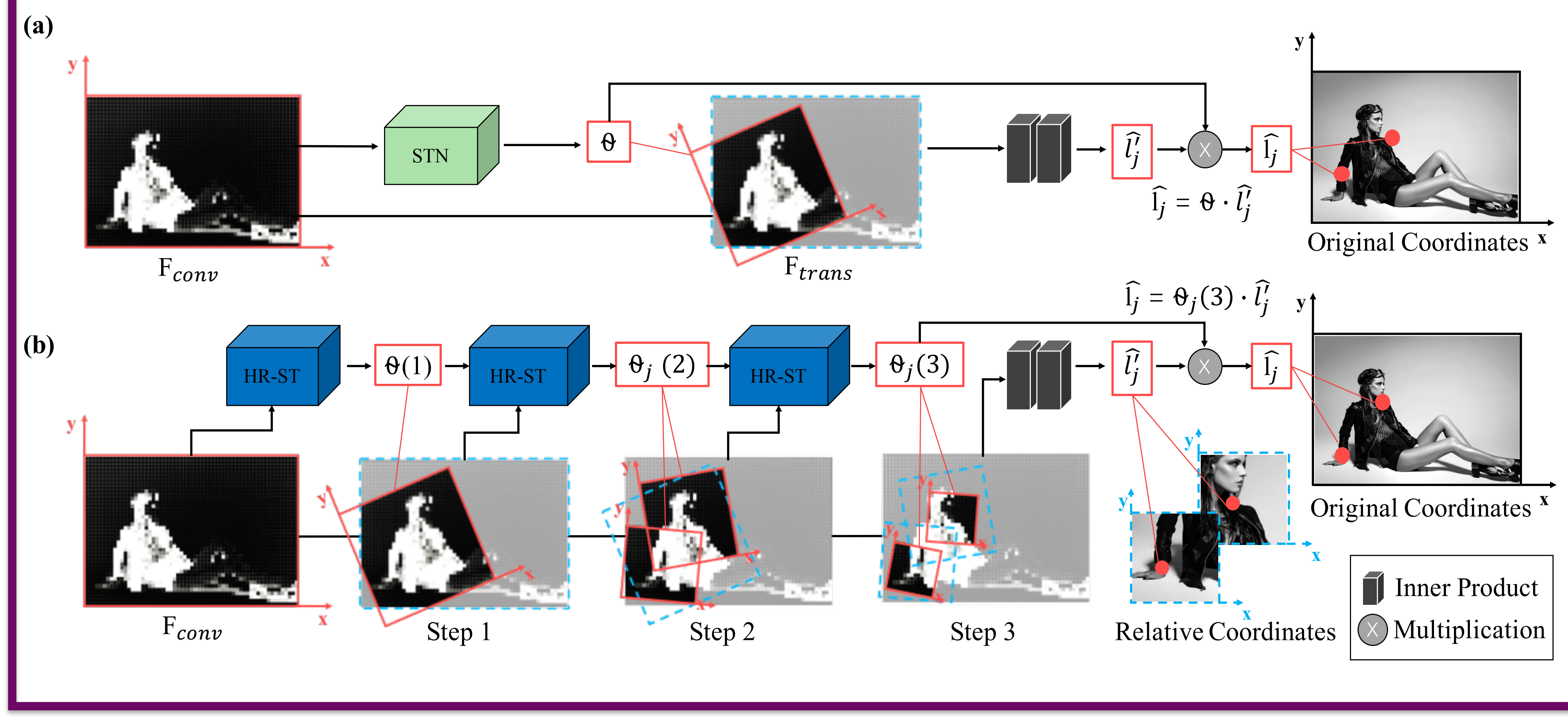
ULD v.s. FLD

- 30K images with comprehensive fashion landmark annotations.
- collected from fashion blogs, forums and online shop.

Approach

Pipeline





Selective Dilated Convolution

- Share weights
- element-wise max

$$F_{conv} = \max_s F_{conv5\star 2^s}$$

F_{conv}: Final convolutional response
s: scale. $F_{conv5\star 2^s}$: convolution
$\star$: expanded sampling.

Hierarchical Recurrent Spatial Transformer

- Recover coordinates $\hat{l}_j = \Theta_{global} \cdot \hat{l_j}'$

$\hat{l_j}'$: relative landmark coordinates
$\hat{l}_j$: original coordinates.
Θ_{global}: geometric transformation for clothes

- Hierarchical model $\Theta_j = \Theta_{global} \cdot \Theta_j$

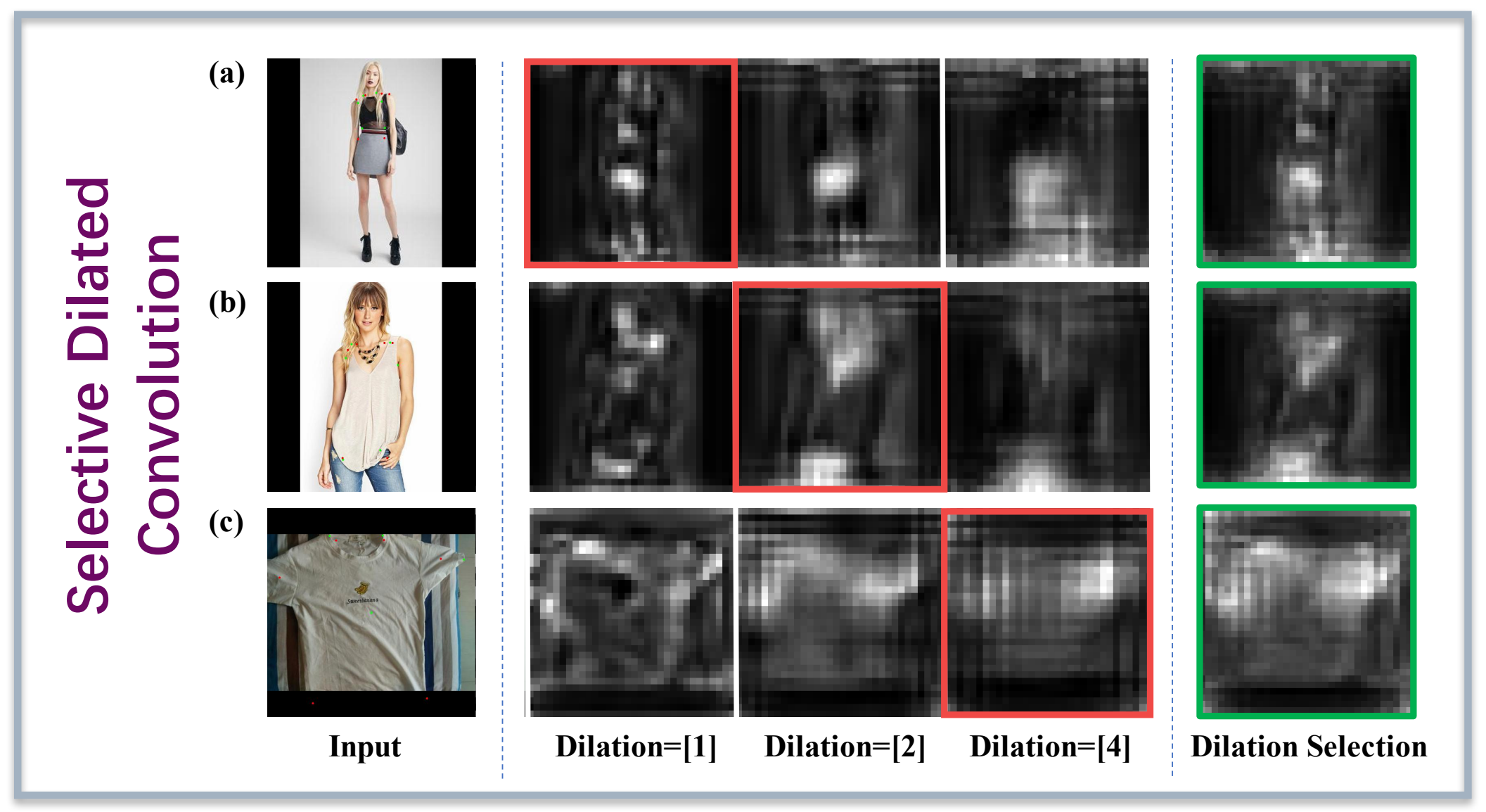
Θ_j: geometric transformation for each landmark
$\Theta_j(i)$: local refinement transformation for recurrent step i

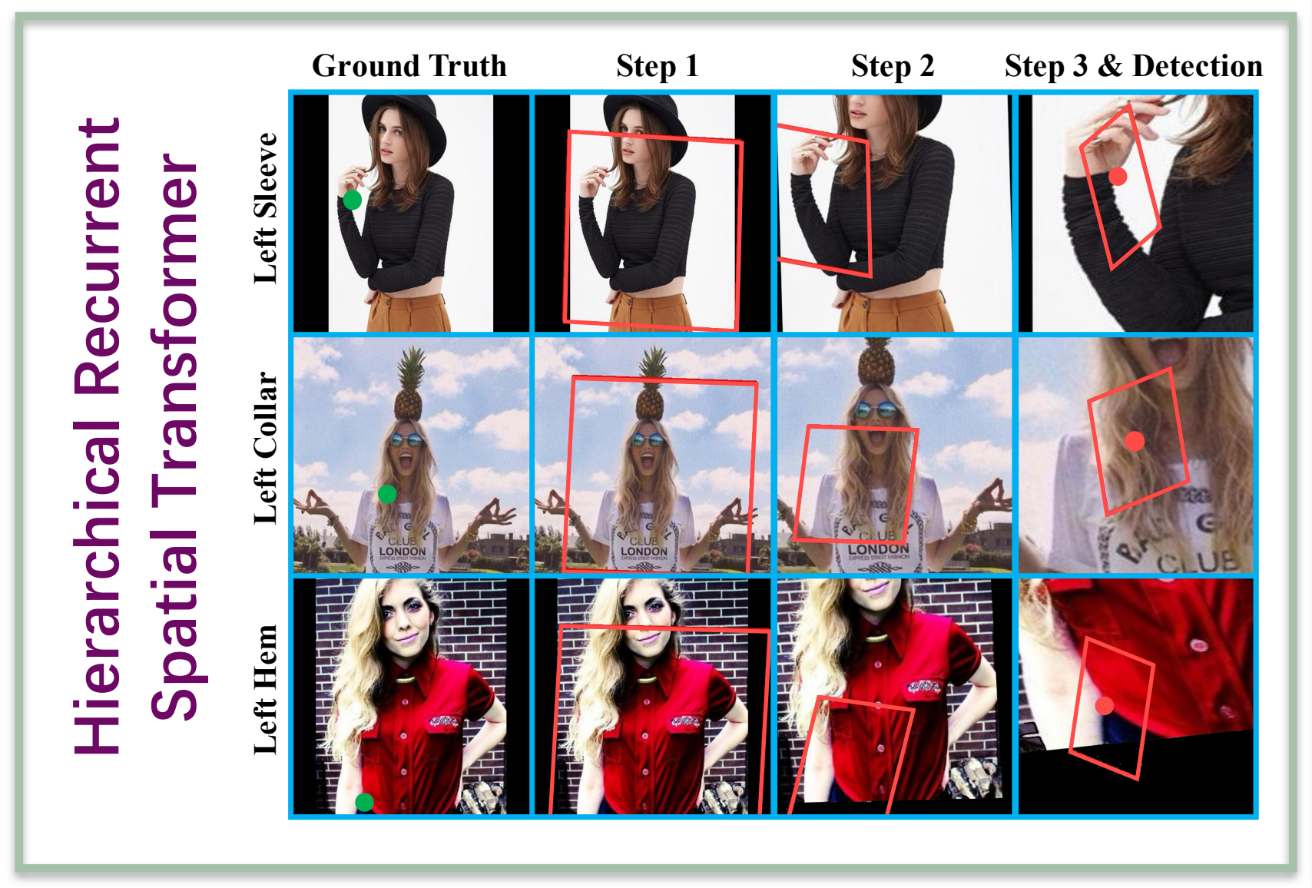
- Recurrent update

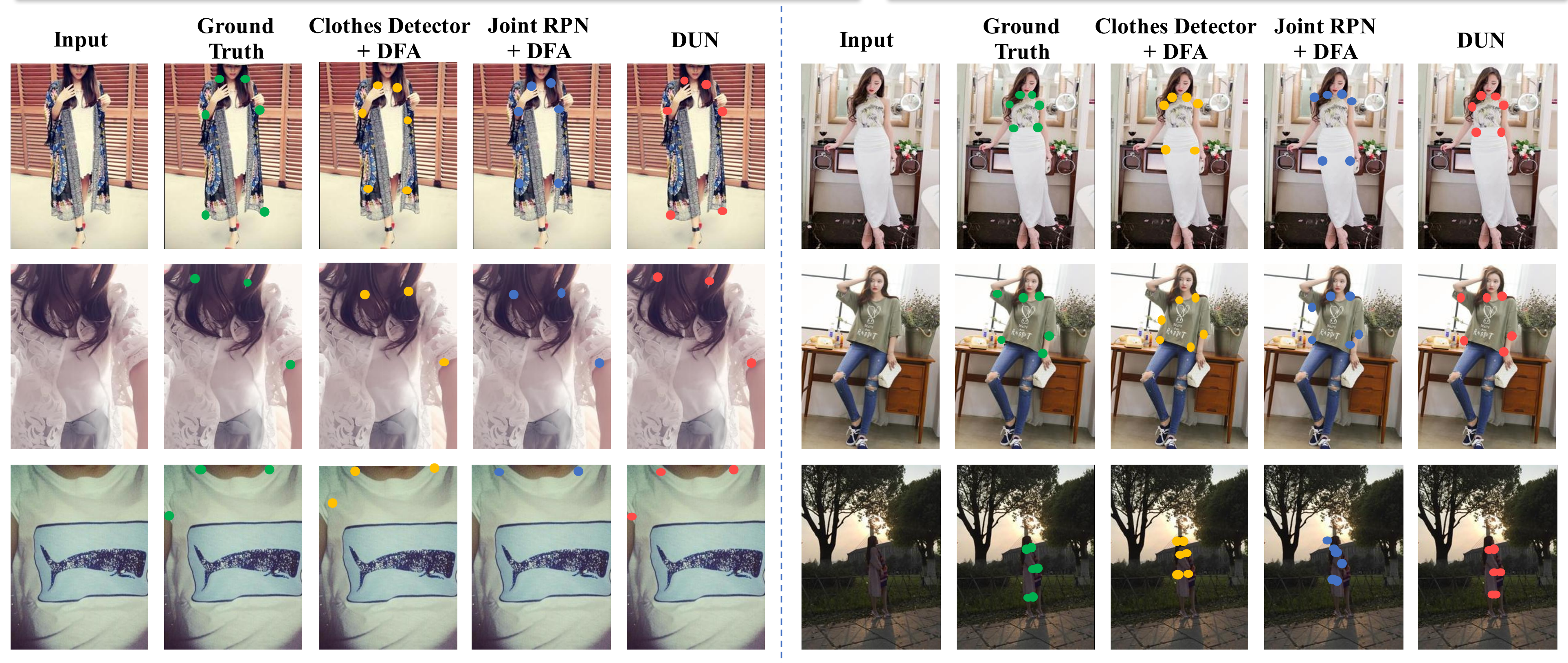
$$\Theta_j(i) \leftarrow \Theta_j(i-1) \cdot \Theta_j(i-1 \to i)$$

$\Theta_j(i-1 \to i)$: refinement transformation

Visual Results







Experiments

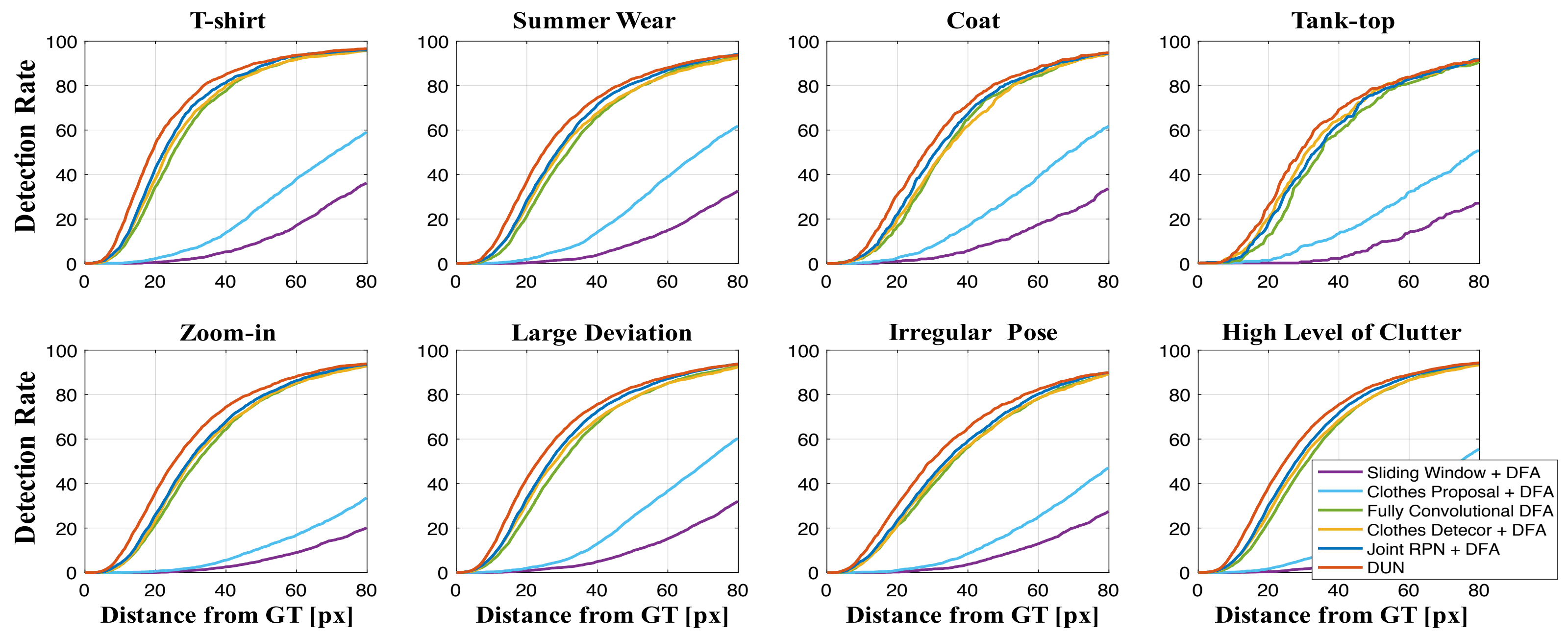
Comparison

	# VGGs	# bbox anno.	end-to-end	# inference pass	speed (fps)	det. rate (%)
Sliding Window + DFA [18]	1	×	×	17	3.2	2.7
Clothes Proposal + DFA [18]	1	×	×	100	0.5	9.7
Clothes Detector + DFA [18]	2	16K	×	1	5.0	63.1
Joint RPN [20] + DFA [18]	2	16K	√	1	3.9	66.0
Deep LAndmark Network	1	×	√	1	**5.2**	**73.8**

Per-landmark Analysis

	L. Collar	R. Collar	L. Sleeve	R. Sleeve	L. Hem	R. Hem	Mean
Fully Convolutional DFA	75.4%	75.7%	52.1%	52.7%	61.2%	61.6%	60.8%
Clothes Detector + DFA	76.3%	76.1%	56.3%	57.6%	61.7%	61.1%	63.1%
Joint RPN + DFA	79.5%	79.8%	55.0%	57.7%	65.4%	66.6%	66.0%
DUN	**83.3%**	**83.7%**	**64.6%**	**66.7%**	**71.7%**	**72.4%**	**73.8%**

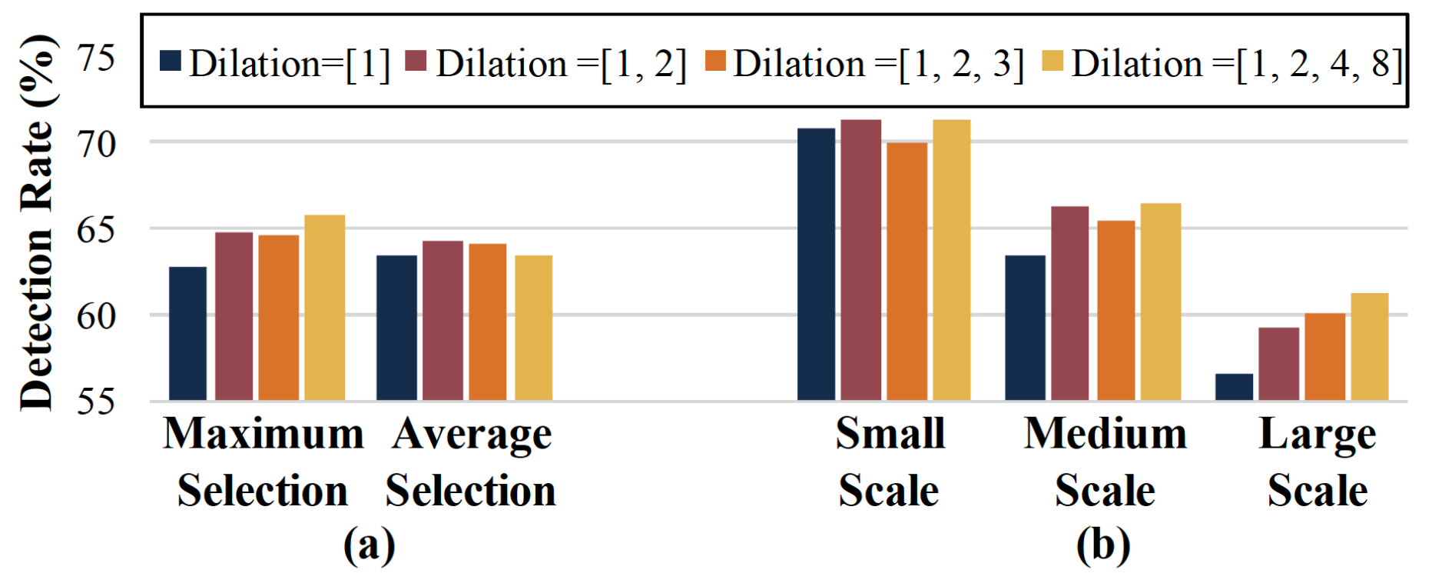
Per-category Analysis



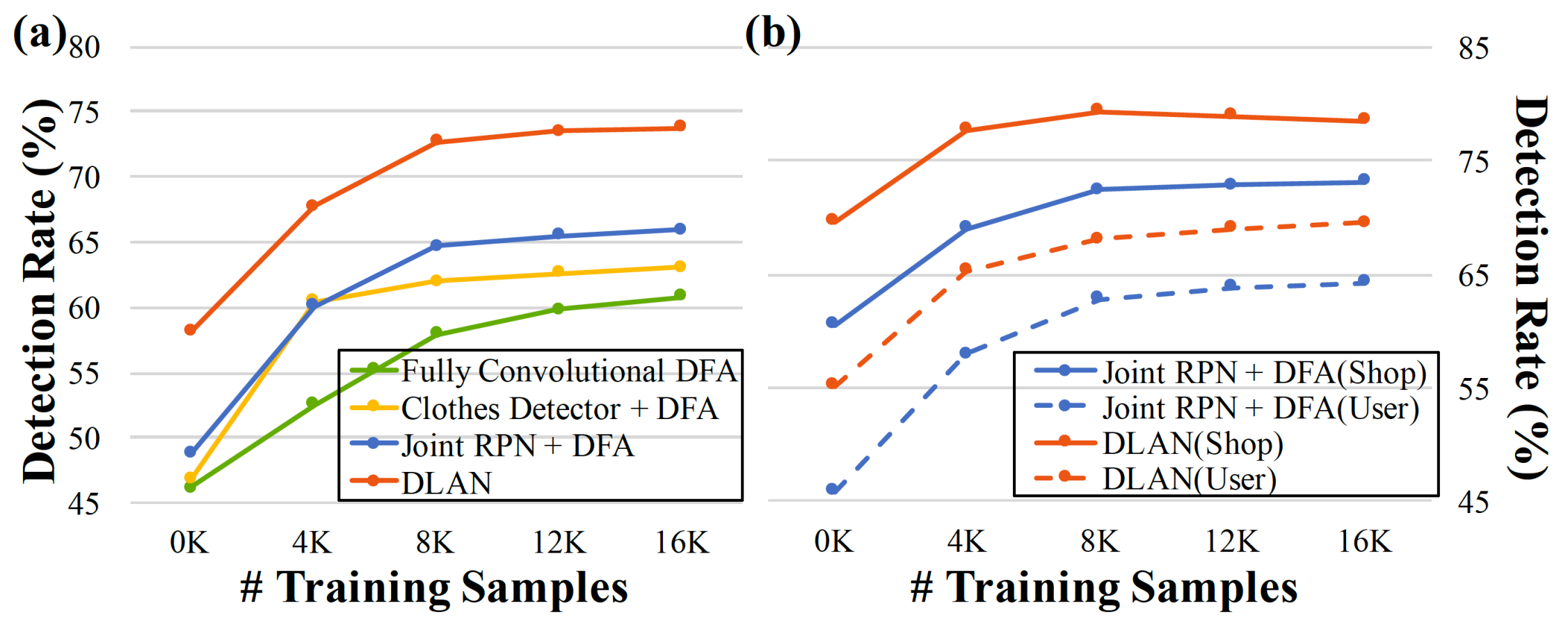
Ablation Study

					DUN
Fully Convolutional DFA?	✓	✓	✓	✓	✓
Spatial Transformer?		✓	✓	✓	✓
Selective Dilated Convolutions?			✓	✓	✓
HR Spatial Transformer?				✓	✓
Scale Regularization?					✓
detection rate (%)	56.9	62.8	64.8	71.2	**73.8**

Selective Dilated Convolution



Generalization of DLAN


(a)
Detection Rate (%)
Fully Convolutional DFA
Clothes Detector + DFA
Joint RPN + DFA
DLAN
0K
4K
8K
12K
16K
# Training Samples
(b)
Joint RPN + DFA(Shop)
Joint RPN + DFA(User)
DLAN(Shop)
DLAN(User)
# Training Samples
Detection Rate (%)