Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization

Chufeng Tang[1], Lu Sheng[2], Zhaoxiang Zhang[3], Xiaolin Hu[1]*

[1]Tsinghua University, [2]Beihang University, [3]Chinese Academy of Sciences

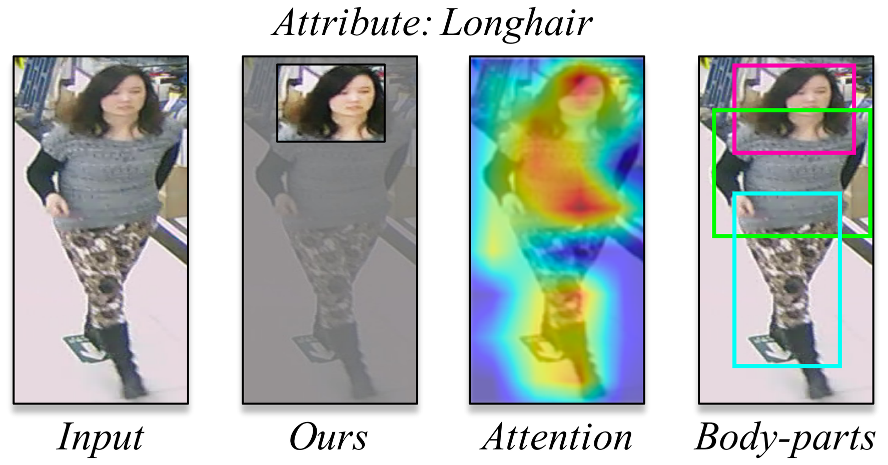
ICCV 2019 Seoul, Korea

Summary

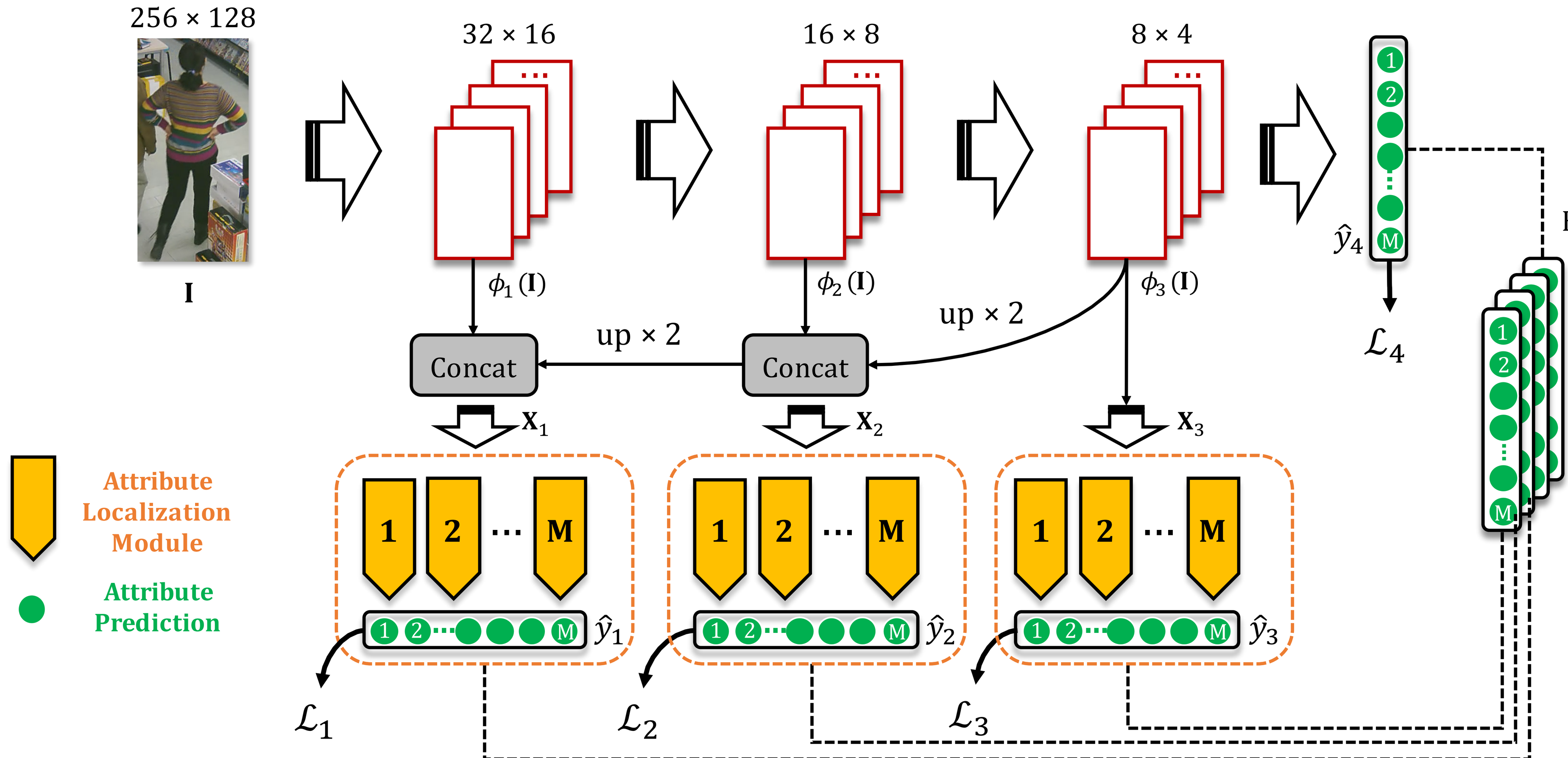
- **Problem**: previous pedestrian attribute recognition methods failed to indicate the attribute-region correspondence
- **Contribution**: performing attribute-specific localization at multiple scales to find the most discriminative region for each attribute in a weakly-supervised manner
- **Results**: improvement across three datasets, end-to-end trainable, less computational cost

Motivation

- *Attribute-agnostic attention*: attend to a broad region, no attribute-region correspondence
- *Rigid body parts localization*: simply fuse the local features, require extra computation
- *We need* ***Attribute-Specific Localization***
  - √ maintain the attribute-region correspondence
  - √ fully adaptive, without region annotations
  - √ interpretable and computationally efficient



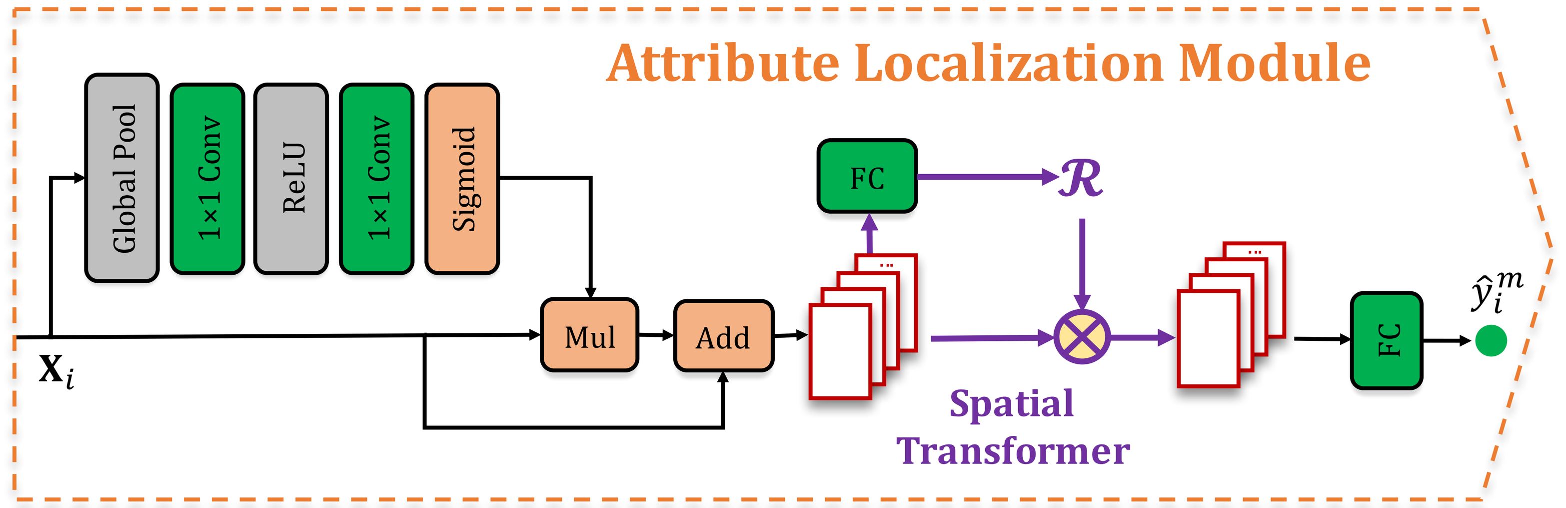
Methodology



- ***Top-down feature pyramid***
  low-level details: feature learning
  high-level semantics: localization
- ***Deep supervision*** *for training*
  4 predictions are directly supervised by GT, trained insufficiently otherwise
- ***Maximum voting*** *for inference*
  choosing the most confident prediction

weighted binary cross-entropy loss $\mathcal{L} = \sum_{i=1}^{4} \mathcal{L}_i$

$$\mathcal{L}_i(\hat{y}_i, y) = -\frac{1}{M}\sum_{m=1}^{M} \gamma^m \left(y^m \log(\sigma(\hat{y}_i^m)) + (1-y^m)\log(1-\sigma(\hat{y}_i^m))\right)$$

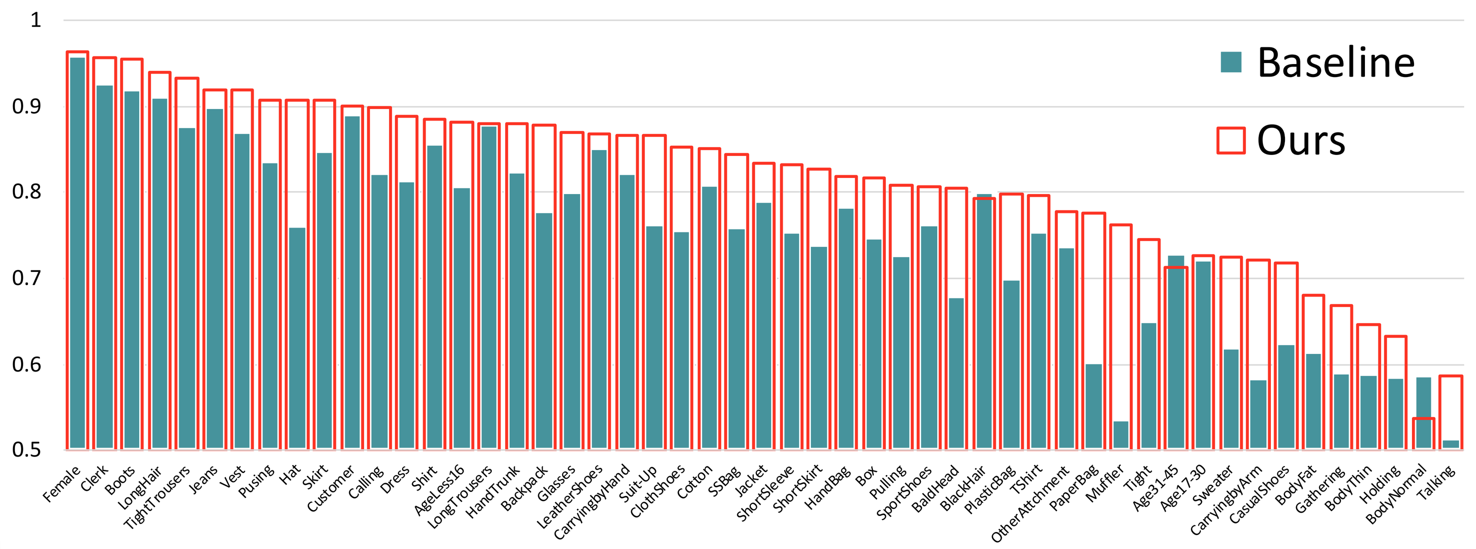


- ***Spatial transformer***
  simplified STN, learn to represent attribute region should be adaptive and differentiable (RoI pooling can't)
- ***Feature alignment***
  a tiny channel attention sub-network, modulating the inter-channel dependencies, since features from different levels should contribute unequally (some need more details).
- ***One for each attribute, but still light-weight***

	DeepMar	PGDM	VeSPA	LG-Net	GRL	BN-Inception	**Ours**
mA	73.79	74.31	77.70	78.68	81.20	75.76	**81.87**
# Params	58.5M	87.2M	17.0M	>20M	>50M	**10.3M**	17.1M
GFLOPs	**0.72**	≈1	> 3	> 4	>10	1.78	1.95

Quantitative Results

Dataset		PETA		RAP		PA-100K	
Method \ Metric		mA	F1	mA	F1	mA	F1
ACN	[ICCVw'15]	81.15	82.64	69.66	75.98	-	-
DeepMar	[ACPR'15]	82.89	83.41	73.79	75.56	72.70	81.32
JRL	[ICCV'17]	85.67	85.42	77.81	78.58	-	-
JRL*	[ICCV'17]	82.13	82.02	74.74	74.62	-	-
GRL	[IJCAI'18]	**86.70**	86.51	81.20	79.29	-	-
HP-Net	[ICCV'17]	81.77	84.07	76.12	78.05	74.21	82.53
VeSPA	[BMVC'17]	83.45	85.49	77.70	79.59	76.32	83.20
DIAA	[ECCV'18]	84.59	86.46	-	-	-	-
PGDM	[ICME'18]	82.97	85.76	74.31	77.35	74.95	83.29
LG-Net	[BMVC'18]	-	-	78.68	80.09	76.96	85.04
BN-Inception		82.66	85.57	75.76	78.20	77.47	85.97
Ours		86.30	**86.85**	**81.87**	**80.16**	**80.68**	**86.46**



Ablation Study

Effectiveness of each component

	Component \ Metric	mA	F1
	Baseline	75.76	78.20
+3.1%	ALM at Single Level (5b)	77.45	79.14
	ALM at Multiple Levels (3b,4d,5b)	78.89	79.50
+1.7%	Top-down (Addition)	78.51	79.42
	Top-down (Concatenation)	79.93	79.91
	Top-down (Channel Attention)	80.61	79.98
+1.3%	Deep Supervision (Averaging)	80.70	80.04
	Deep Supervision (Maximum) (**Ours**)	**81.87**	**80.16**
	Ours w/o ALMs	78.91	79.55

Three different attribute-specific methods

- Each ***attention mask*** corresponds to one attribute
  over-adaptive: try to cover all pixels but often failed, since there is no accurate localization labels.
- Each attribute associated with ***predefined parts***
  lack-adaptive: discard the adaptive factors, which are less robust to variances.
- We ***achieve a balance***
  between two extremes using attribute-specific bounding boxes, which relatively coarse but more interpretable.

Method \ Metric	mA	F1
Rigid Part	76.56	78.84
Attention Mask	78.35	79.51
Attribute Region	**81.87**	**80.16**