Attribute Recognition from Adaptive Parts

An end-to-end learning approach for localized attribute recognition

Luwei Yang[1], Ligeng Zhu[12], Yichen Wei[3], Shuang Liang[4], Ping Tan[1]

[1]*Simon Fraser University*, [2]*Zhejiang University*

[3]*Microsoft Research Asia*, [4]*Tongji University*

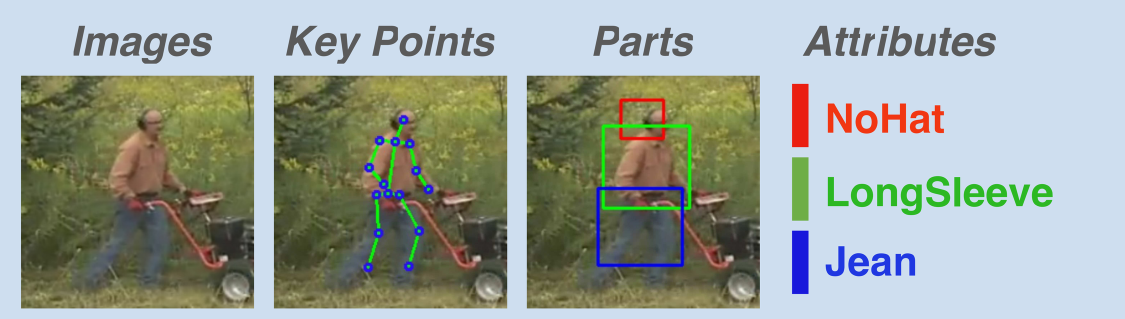
gruvi {graphics} ∪ {vision}

SFU SIMON FRASER UNIVERSITY

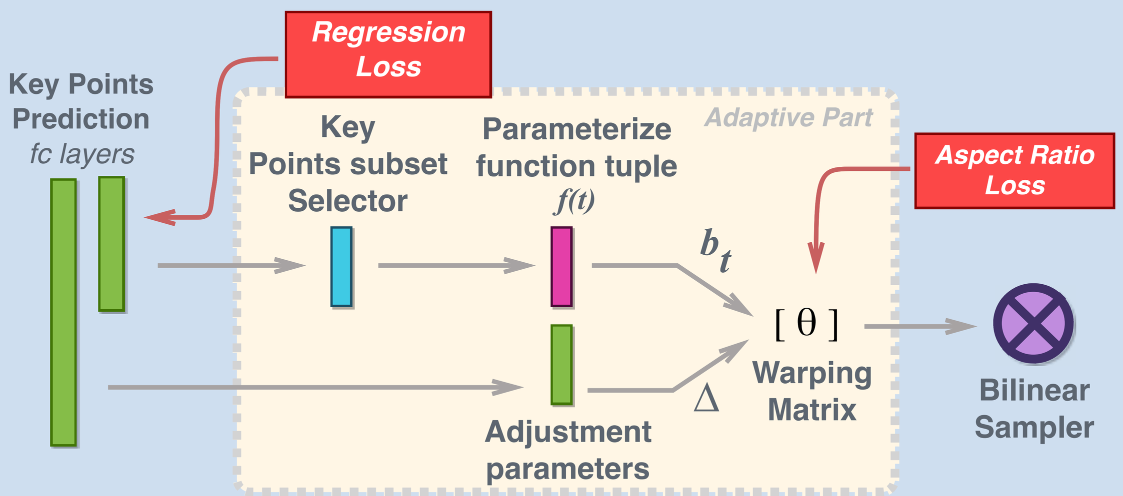
Introduction

- Attribute recognition is usually treated as classification of the whole object. This is undesirable for localized attributes where only local regions are useful.
- Works in [1, 2] show part-based approach has better performance, but they addressed the problem as two-step approach: parts are firstly detected and then used for attribute recognition.
- Inspired by the recent spatial transformer network [3], we proposed an end-to-end deep learning approach to optimize part detection for attribute recognition.

Overview



Adaptive parts



The **Adaptive Part** is responsible for learning the part localization for a certain attribute:

- Initial bounding box: $b_t = [w_t, h_t, x_t, y_t]$
- Learnable adjustment: $\Delta = [\Delta_w, \Delta_h, \Delta_x, \Delta_y]$
- The final bounding box is encoded by the wrapping matrix:

$$\theta_t = \begin{bmatrix} w_t(1+\Delta_w) & 0 & x_t+\Delta_x \\ 0 & h_t(1+\Delta_h) & y_t+\Delta_y \end{bmatrix}. \quad (1)$$

The **Aspect Ratio Loss** is introduced to restrict the aspect ratio of bounding box:

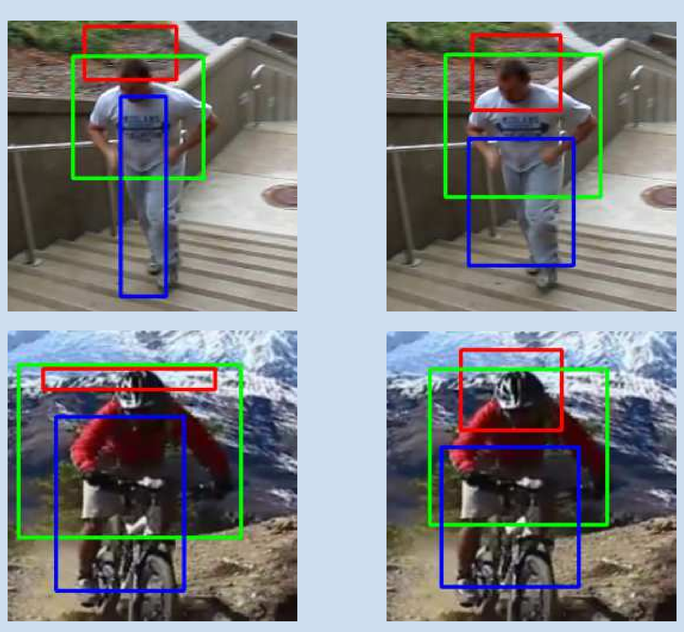
- if $h_t(1+\Delta_h) > w_t(1+\Delta_w)$:

$$L_r^t = \frac{1}{2}\{[\alpha[h_t(1+\Delta_h)]^2 - [w_t(1+\Delta_w)]^2\}_+ \quad (2)$$

- if $w_t(1+\Delta_w) > h_t(1+\Delta_h)$:

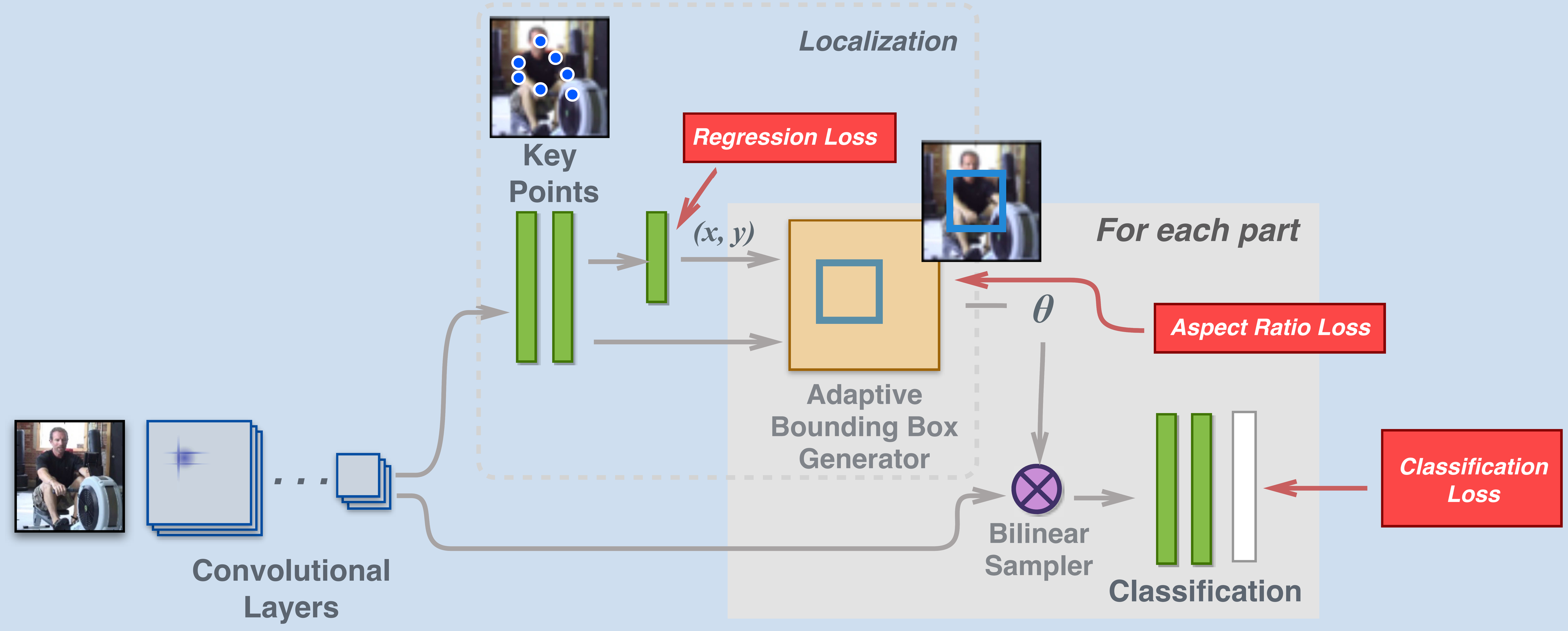
$$L_r^t = \frac{1}{2}\{[\alpha[w_t(1+\Delta_w)]^2 - [h_t(1+\Delta_h)]^2\}_+ \quad (3)$$

- Examples:

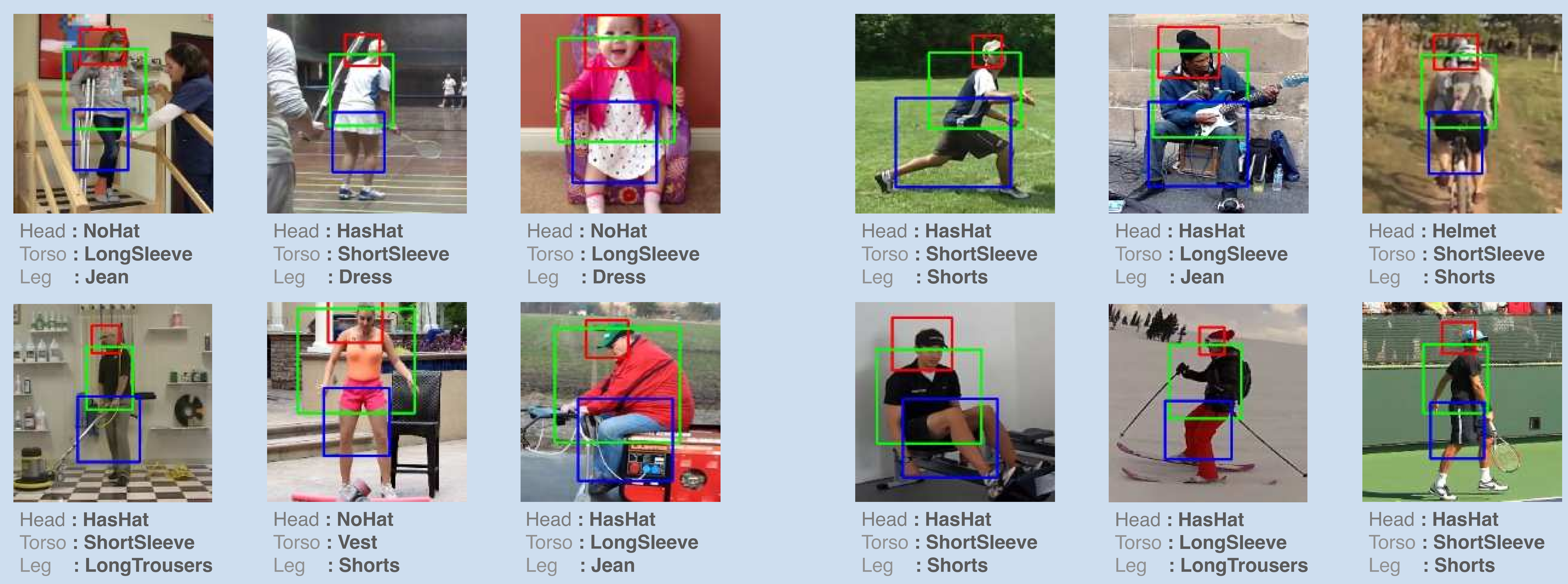


RatioLoss(×) RatioLoss(√)

Network Structure



Example Results



MPII Dataset

Experiment on Augmented MPII

Table below is based on augmented MPII with VGG-16.

- **Full**: The pipeline uses whole image as input and is considered as *lower bound.*
- **STN**: Single spatial transformer module without feedback.
- **Separate**: localize the key points first, then do attribute recognition separately.
- **Ours**: Adaptive parts detector initialized by key points.
- **Oracle**: Similar to **Separate**, but it uses ground-truth key points directly. This is taken as *upper bound* of all methods.

Attribute/Pipeline	Full	STN	Separate	Ours	Oracle
Helmet	81.76	80.58	57.60	**83.53**	84.04
HasHat	79.16	78.18	57.51	**81.21**	83.75
NoHat	96.39	**96.78**	88.30	96.45	97.15
Avg. Accuracy	84.25	83.96	74.00	**86.02**	86.16
LongSleeve	83.62	84.22	83.51	**87.89**	88.52
Vest	80.38	80.64	80.86	**81.57**	83.49
ShortSleeve	88.99	88.67	88.43	**91.35**	92.76
Naked	49.73	54.99	47.43	**61.18**	49.41
Avg. Accuracy	73.60	75.68	74.51	**79.68**	78.94
Jean	67.75	69.18	67.31	**69.58**	73.55
Dress	26.30	34.81	20.98	**38.45**	37.81
Shorts	91.46	91.21	89.42	**93.42**	92.97
Trousers	89.04	89.51	86.96	**90.83**	90.90
Avg. Accuracy	79.00	80.82	78.71	**82.22**	82.25

References

[1] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[2] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.