Diversity in Fashion Recommendation Using Semantic Parsing

Sagar Verma[1], Sukhad Anand[1], Chetan Arora[1], Atul Rai[2]

[1]Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi.

[2]Staqu Technologies

ICIP, 2018

Problem Statement

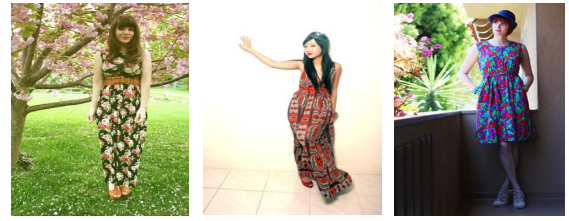
Recommendation based on contextual similarity

Images retrieved by finding similarity between features computed over whole image.

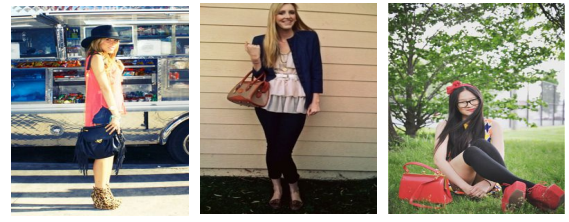
Images retrieved by finding similarity between features computed over semantically similar parts of image.

Query Image

Hat



Dress



Bag

Relevant Item Images

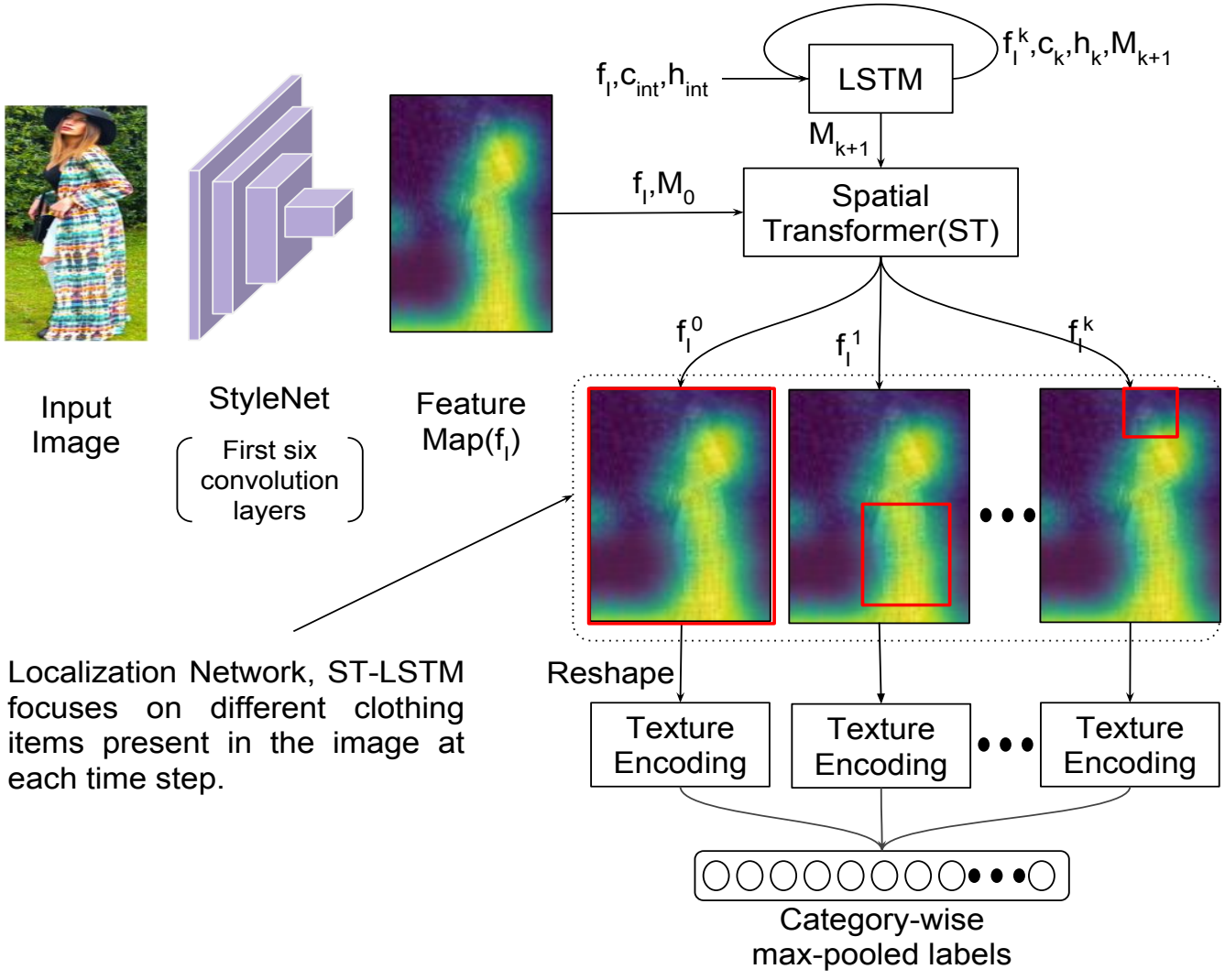
Contributions

1. Our method recommends similar images based on different parts of a query image.
2. To identify different parts we use attention and weakly labeled data.
3. Instead of features from standard pre-trained neural networks, we suggest using texture-based features.
4. We evaluate our method on item recognition task, consumer-to-shop retrieval and in-shop retrieval tasks.

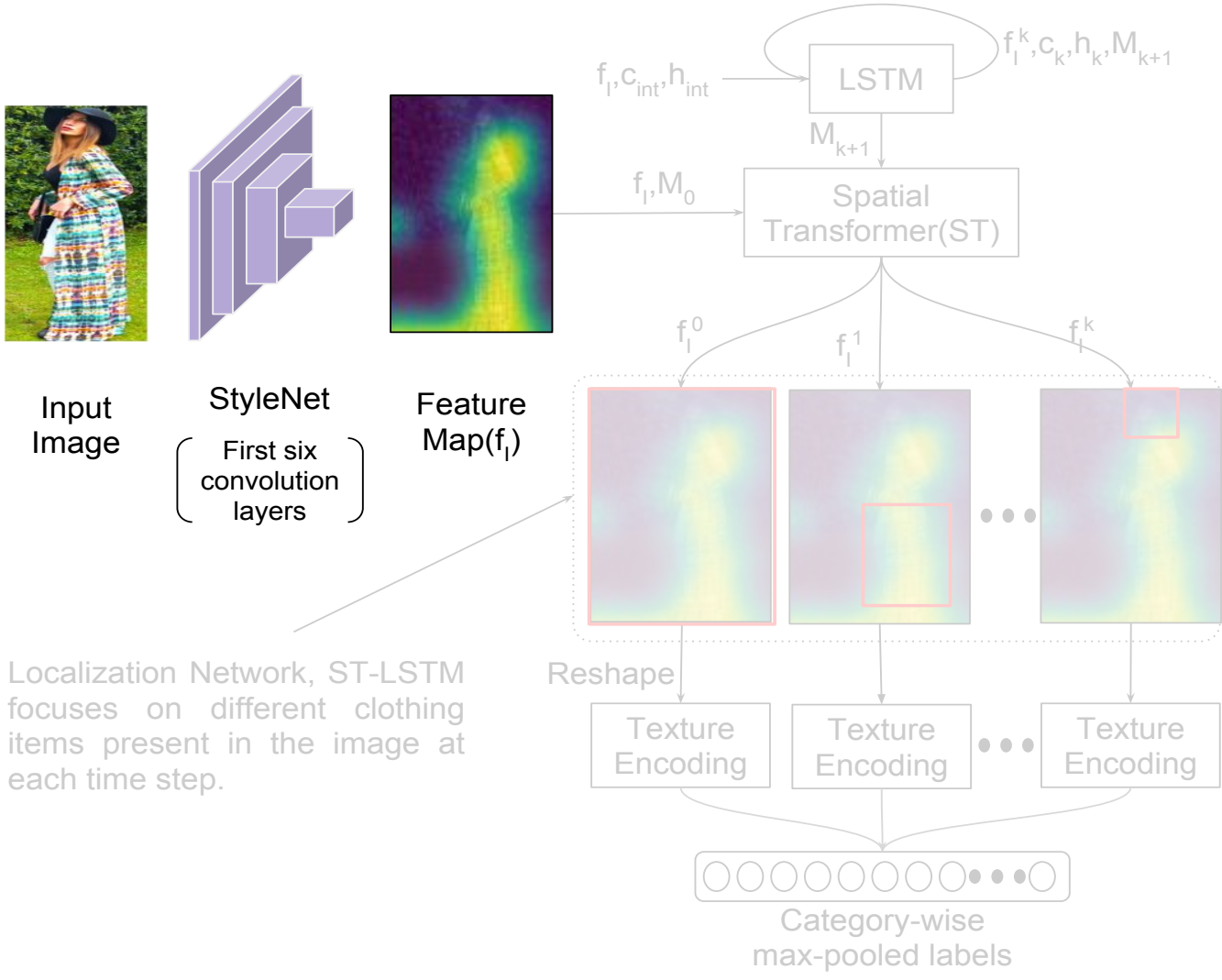
Related Work

- Cloth parsing,
- Clothing attribute recognition,
- Detecting fashion style, and
- Cross-domain image retrieval using Siamese network and Triplet network.

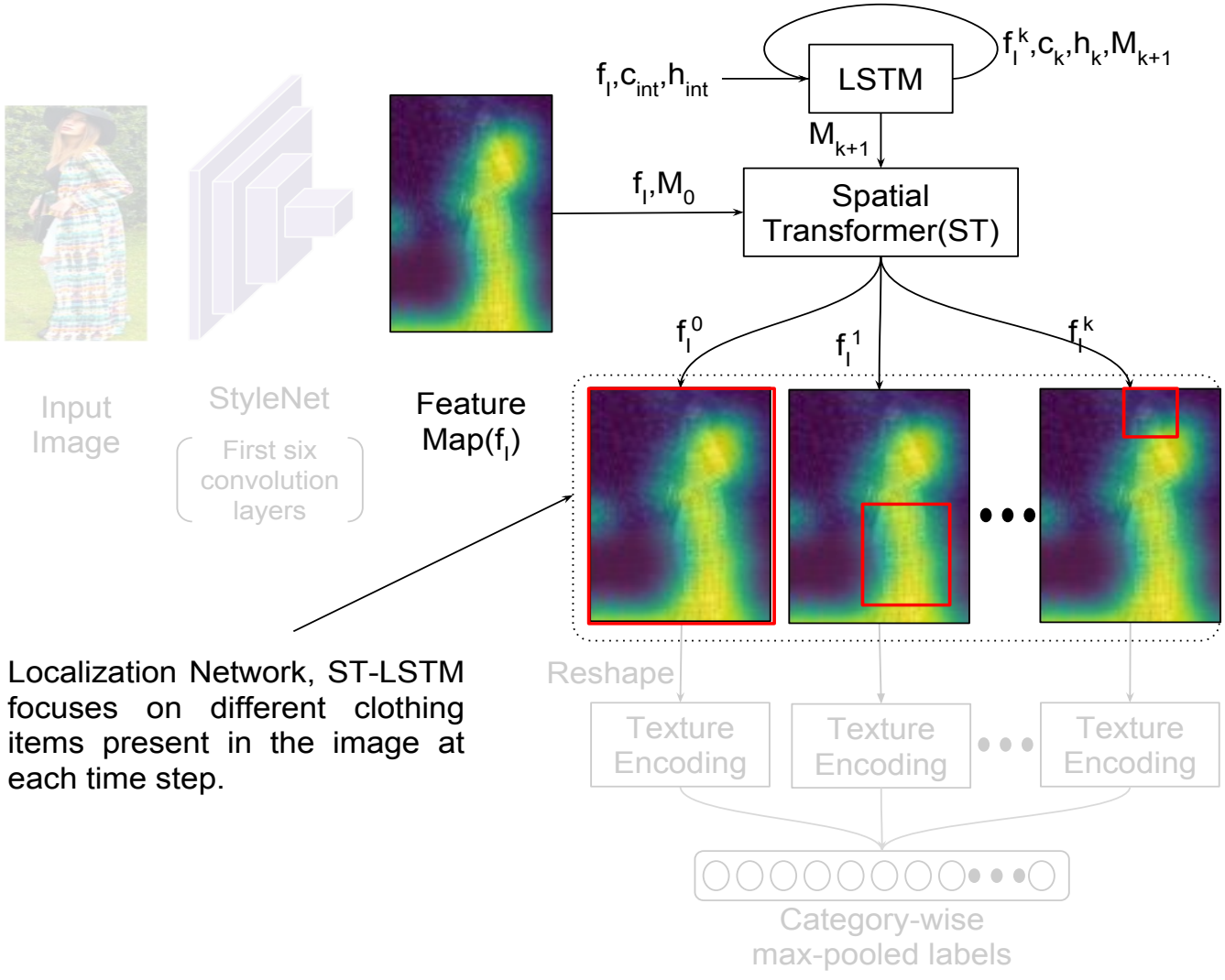
Proposed Architecture



CNN for Global Image Features

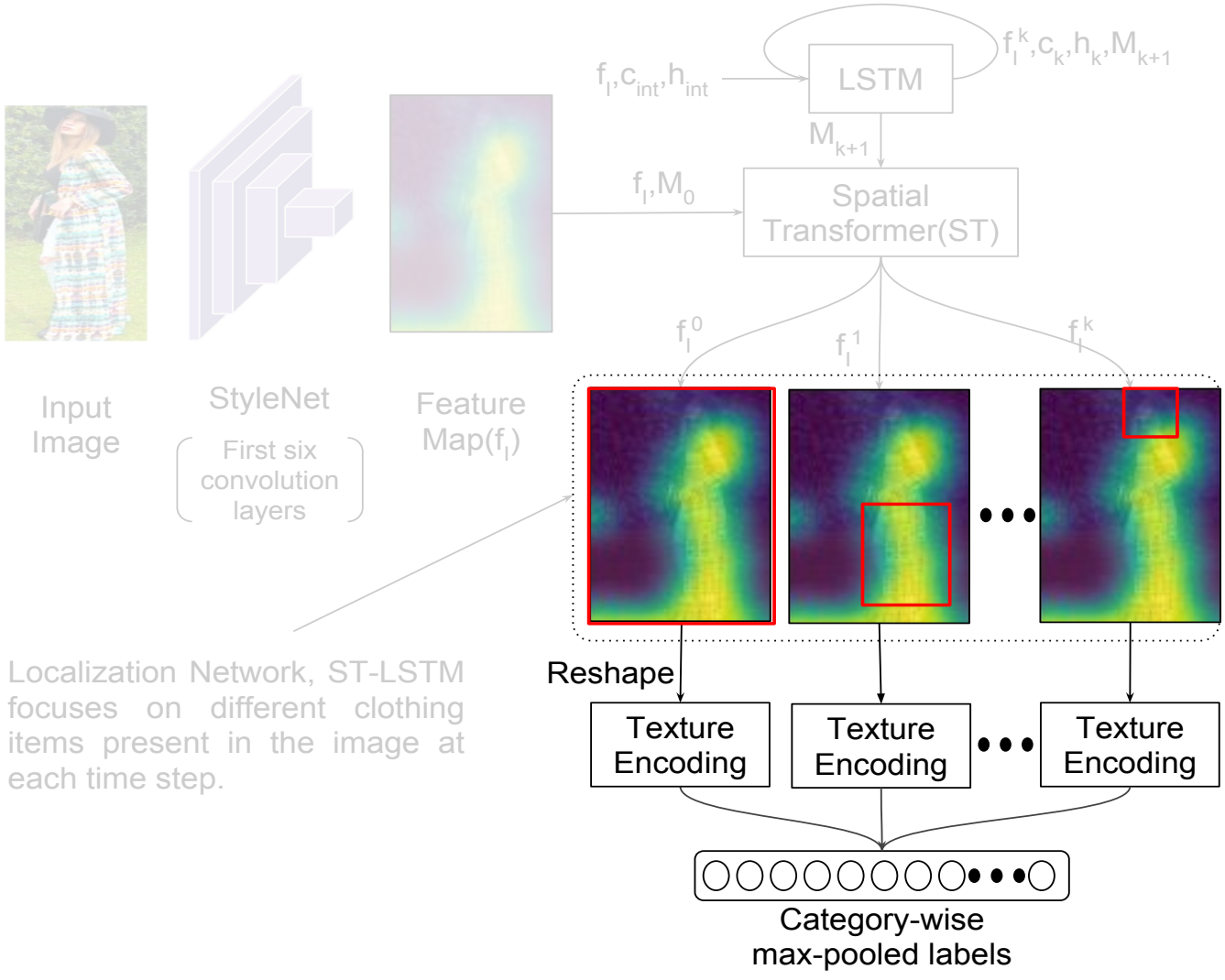


Visual Attention Module



Localization Network, ST-LSTM focuses on different clothing items present in the image at each time step.

Texture Encoding Layer



Multi-label classification loss

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} (p_i^c - \hat{p}_i^c)^2 \tag{1}$$

where, N is training sample, C is total number of classes, $\hat{p}_i$ is ground truth label vector of sample i and p_i is predicted label vector of sample i.

Diversity loss

Diversity loss is the correlation between adjacent attention maps,

$$\mathcal{L}_{div} = \frac{1}{K-1} \sum_{k=2}^{K} \sum_{i=1}^{HxW} l_{k-1,i} . l_{k,i} \tag{2}$$

where, K is the total steps of recurrent attention, HxW is the height and width of attention maps, l_k is the k^{th} attention map.

Localisation loss

Localization loss, $\mathcal{L}_{loc}$ from [1] is used to remove redundant locations and force localization network to look at small clothing parts.

Combined Loss

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{div} + \lambda_2 \mathcal{L}_{loc} \tag{3}$$

where λ_1 and λ_2 are multiplicative factors. We use 0.01 for all our experiments.

Datasets

- **Fashion144K [2]**
  - 90,000 images with multilabel annotation.
  - 128 classes.
  - Image resolution is $384x256$.
- **Fashion550K [3]**
  - 66 classes.
- **DeepFashion [4]**
  - 800,000 images
  - Similarity pairs is available for consumer-to-shop and in-shop retrieval tasks

Experiments

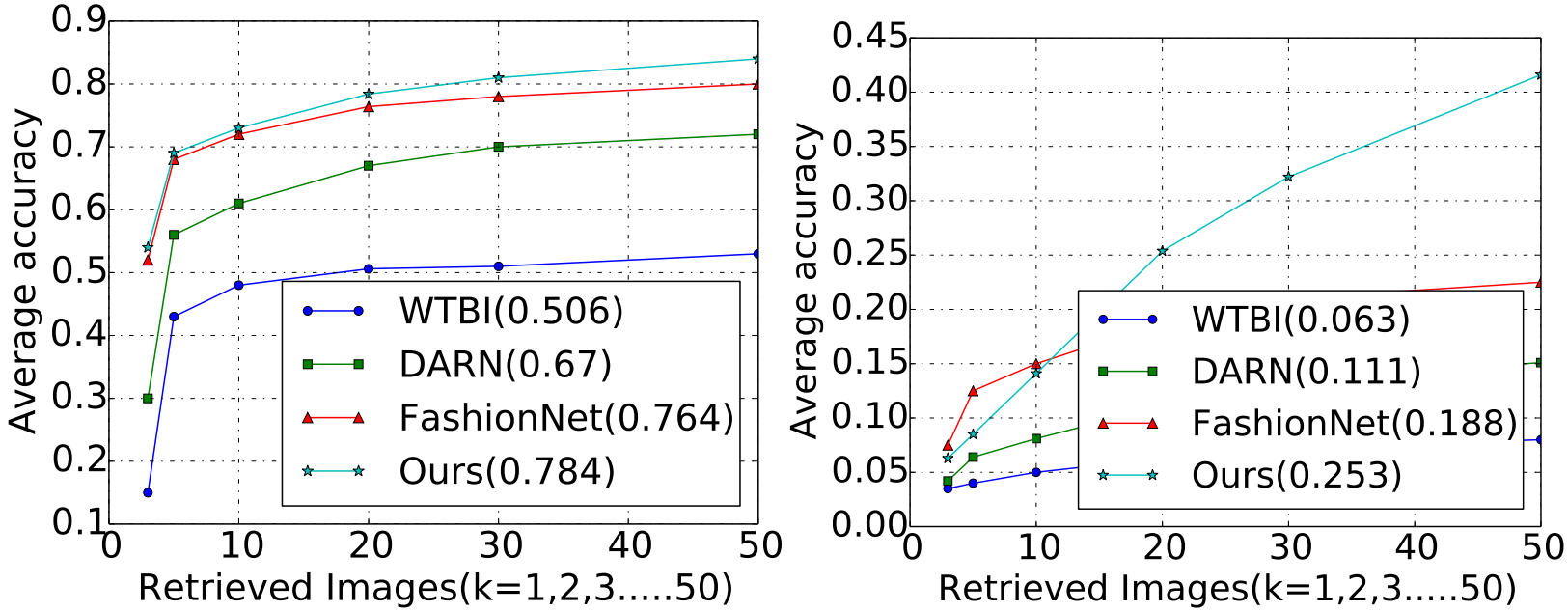
- Model is trained on Fashion144K [2] dataset with 59 item labels, color labels were excluded.
- Evaluated item recognition task on Fashion144K [2] and Fashion550K [3] dataset.
- Consumer-to-shop and in-shop retrieval tasks are evaluated on DeepFashion [4] dataset.

Results

Dataset	Fashion144k [2]		Fashion550k [3]	
Model	AP_{all}	mAP	AP_{all}	mAP
StyleNet [2]	65.6	48.34	69.53	53.24
Baseline [3]	62.23	49.66	69.18	58.68
Viet et al. [5]	NA	NA	78.92	63.08
Inoue et al. [3]	NA	NA	79.87	**64.62**
Ours	**82.78**	**68.38**	**82.81**	57.93

Multi-label classification on Fashion144k [2] and Fashion550k [3]

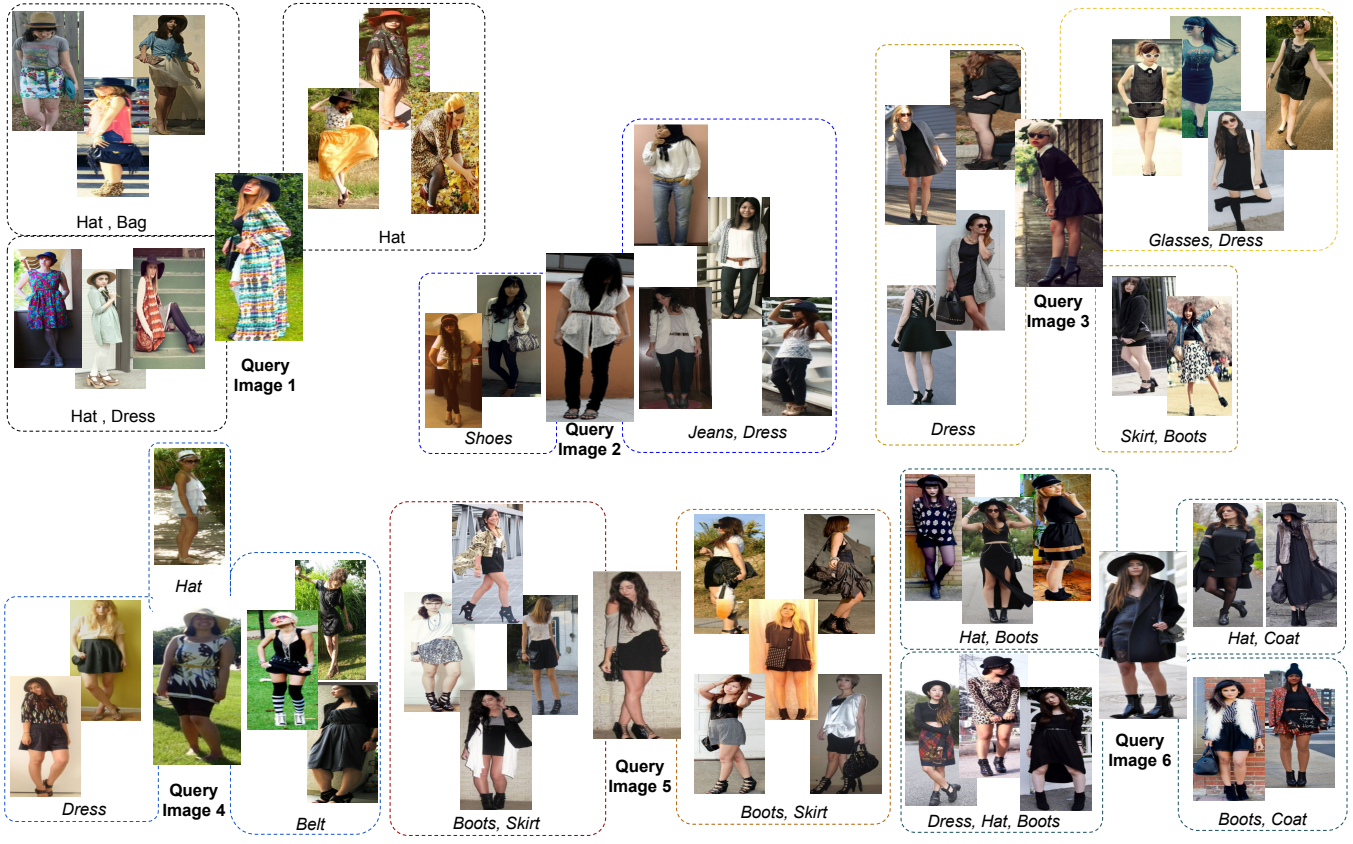
Results



(a) In-Shop retrieval (b) Consumer-to-shop retrieval

Retrieval results for In-shop and Consumer-to-shop retrieval tasks on DeepFashion dataset [4].

Results



Semantically similar results for some of the query images from Fashion144k dataset [2] using our method.

Conclusion

- Using clothing parts for recommendation gives much variability in the recommendation results.
- Attention can be used to learn discriminative features from weak labels.
- Texture cues are important for learning different parts.

References

Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin,
"Multi-label image recognition by recurrently discovering attentional regions,"
in *ICCV*, 2017.

E. Simo-Serra and H. Ishikawa,
"Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction,"
in *CVPR*, 2016, pp. 298–307.

Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa,
"Multi-label fashion image classification with minimal human supervision,"
in *ICCVW*, 2017, pp. 2261–2267.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang,
"Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,"
in *CVPR*, 2016, pp. 1096–1104.

Andreas Veit et al.,
"Learning from noisy large-scale datasets with minimal supervision,"
in *CVPR*, 2017.

Thank you
Questions?