DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations

Supplementary Material

Ziwei Liu[1] Ping Luo[3,1] Shi Qiu[2] Xiaogang Wang[1,3] Xiaoou Tang[1,3]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited [3]Shenzhen Institutes of Advanced Technology, CAS
{lz013,pluo,xtang}@ie.cuhk.edu.hk, sqiu@sensetime.com, xgwang@ee.cuhk.edu.hk

1. Labels in DeepFashion Dataset

To illustrate the labels in DeepFashion dataset, the 50 fine-grained fashion categories and massive fashion attributes are listed in Table 1 and 2, respectively. As described in paper line 358 $\sim$ 365, we also define a set of clothing landmarks, which corresponds to a set of key-points on the structures of clothes. The detailed clothing landmark definitions for upper-body clothes, lower-body clothes and full-body clothes are listed in Table 3.

Upper-cloth (20)	Anorak, Blazer, Blouse, Bomber, Button-Down, Cardigan, Flannel, Halter, Henley, Hoodie, Jacket, Jersey, Parka, Peacoat, Poncho, Sweater, Tank, Tee, Top, Turtleneck
Lower-cloth (14)	Capris, Chinos, Culottes, Cutoffs, Gauchos, Jeans, Jeggings, Jodhpurs, Joggers, Leggings, Sarong, Shorts, Skirt, Sweatshorts, Trunks
Full-cloth (16)	Caftan, Cape, Coat, Coverup, Dress, Jumpsuit, Kaftan, Kimono, Nightdress, Onesie, Robe, Romper, Shirtdress, Sundress

Table 1: List of fine-grained fashion categories.

Texture	Baroque, Butterfly, Brocade, Chevron, Clean, Colorblock, Contrast, Daisy, Diamond, Dot, Distressed, Embellished, Floral, Frond, Geo-Patterned, Grid, Houndstooth, Kladoscope, Leopard, Mandala, Marble, Mineral, Mosaic, Paisley, Palm, Panel, Pinstriped, Plaid, Raglan, Ringer, Southwestern-Print, Speckled, Splatter, Star, Stripe, Tartan, Tile, Triangle, Two-Tone, Watercolor, Windowpane, . . .
Fabric	Chiffon, Chino, Cotton, Denim, Damask, Dip-Dye, Embroidered-Mesh, Frayed, Fur, Heather, Lace, Leather, Linen, Loose-Knit, Metallic, Open-Knit, Organza, Pleated, Pointelle, Quilted, Ribbed, Satin, Sequined, Shaggy, Sleek, Slub, Stretch-Knit, Suede, Thermal, Tie-Dye, Tulle, Tweed, Twill, Velveteen, Waffle, Washed, Woven, . . .
Shape	A-Line, Boxy, Batwing, Crop, Fit, High-Rise, Layered, Longline, Low-Rise, Maxi, Mid-Rise, Midi, Mini, Oversized, Pencil, Popover, Sheath, Skinny, Slim, Slouchy, Smock, Tiered, Trapeze, Tube, Tunic, Vented, . . .
Part	Bow-Front, Button-Front, Cap-Sleeve, Collar, Collarless, Crew-Neck, Crochet-Trimmed, Crisscross-Back, Cuffed-Sleeve, Cutout-Back, Double-Breasted, Drop-Sleeve, Flared, Flounce, Fringed, High-Low, High-Neck, High-Slit, Hooded, Keyhole, Knotted, Ladder-Back, Long-Sleeved, M-Slit, Off-The-Shoulder, Open-Shoulder, Peplum, Pintucked, Pocket, Racerback, Ruffled, Shoulder-Strap, Side-Cutout, Single-Button, Sleeveless, Split-Neck, Strappy, Tasseled, Tie-Front, Topstitched, Tulip-Back, Twist-Front, V-Back, V-Cut, V-Neck, Y-Back, Zip-Up, . . .
Style	Baseball, Beach, Boyfriend, Brooklyn, Cargo, Chic, Folk, Graphic, Mickey, Muscle, Nautical, Ornate, Peasant, Polka, Relaxed, Regime, Retro, Rugby, Sky, SpongeBob, Sweetheart, Surplice, Tribal, Trench, Varsity, Wild, Workout, Yoga, . . .

Table 2: List of massive fashion attributes.

2. Data Quality

We have taken into consideration the quality of labelling when using meta-data to generate clothing attributes. We discarded images with too few textual meta-data. After automatically annotating attributes, human annotators also conducted a fast screening to rule out falsely 'fired' images for each attribute to ensure the precision of the attribute labels. For other manually annotated labels, we collected annotations from two different annotators and check their consistency. Around 0.5% samples were found inconsistent and required further labelling from a third annotator.

Admittedly a considerable portion of positive samples have been falsely annotated as negatives for an attribute. However, the accuracy of negative annotations remains high, as the total number of images in the database is huge with most of which being true negatives. For a quantitative assessment, we sample a subset of 100 attributes and manually grade 500 'fired' and 500 'unfired' images per attribute, as has been done in [2]. We find

Upper-body Clothes (6)	Left Collar End, Right Collar End, Left Sleeve End, Right Sleeve End, Left Hem, Right Hem
Lower-body Clothes (4)	Left Waistline, Right Waistline, Left Hem, Right Hem
Full-body Clothes (8)	Left Collar End, Right Collar End, Left Sleeve End, Right Sleeve End, Left Waistline, Right Waistline, Left Hem, Right Hem

Table 3: Clothing landmark definitions for upper-body clothes, lower-body clothes and full-body clothes, respectively.

the accuracies for positive and negative annotations (*i.e.* $\frac{\sum True\,positive}{\sum Annotated\,positive}$ and $\frac{\sum True\,negative}{\sum Annotated\,negative}$) are 97.0% and 99.4%, respectively. Therefore, our attribute labels can serve as an effective training source.

3. Network Architecture of FashionNet

FashionNet employs VGG-16 [3] as backbone, as indicated in paper line 455 $\sim$ 462. Here, we illustrate the detailed pipeline of FashionNet in Fig.1, the network architecture (including network configuration and hyper-parameters) of which is listed in Table 4.

4. Additional Experiments on FashionNet

We conducted additional experiments on the in-shop clothes retrieval benchmark and reported the top-20 retrieval accuracies.

Ablation Study We remove rich attribute supervision, landmark prediction/pooling, and triplet loss incorporating pair correspondences from our full model, respectively. The results in Table 5 (a) suggest that all components in FashionNet are beneficial and complementary.

Attribute Selection To assess the importance of different attribute groups, we equip FashionNet with 100 attributes from each group and compare their performance. Table 5 (b) illustrates that "texture" and "part" attributes are more useful to capture discriminative traits.

Combining Landmarks Table 5 (c) shows that combining human joints, poselets and fashion landmarks only leads to marginal gain. Therefore, fashion landmarks are effective and sufficient local representation for clothes.

Results of Landmark Visibility Prediction Besides fashion landmark locations, FashionNet also predicts landmark visibility to gate local features. In this section, we present the results of landmark visibility prediction. Table 5 (d) provides the visibility prediction accuracy for each clothing landmark. We observe FashionNet achieves nearly 90% visibility prediction accuraries for all clothing landmarks. Sleeve landmarks have relatively low accuraries because of the frequent occlusions introduced by hair.

w/o attr	w/o landmarks	w/o pair	full model
54.3%	66.2%	46.5%	76.4%

(a) Performance of removing different components.

Texture	Fabric	Shape	Part	Style
59.3%	54.6%	57.1%	60.2%	54.9%

(b) Performance of using different attribute groups.

human joints	poselets	fashion landmarks	combined
68.2%	69.9%	76.4%	77.3%

(c) Performance of combining landmarks.

Left Collar.	Left Sleeve.	Left Waistline	Left Hem
87.12%	93.67%	92.46%	94.73%
Right Collar.	Right Sleeve.	Right Waistline	Right Hem
88.46%	93.94%	92.71%	95.17%

(d) Landmark visibility prediction accuracy for each clothing landmark.

Table 5: Additional experimental results of FashionNet. The top-20 retrieval accuracies on the in-shop clothes retrieval benchmark are reported.

5. More Visual Results of Clothes Retrieval

Fig.2 and Fig.3 demonstrate more visual results on in-shop clothes retrieval benchmark and consumer-to-shop clothes retrieval benchmark, respectively. FashionNet is capable of handling complex variations in both scenarios.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 3

[2] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 108(1-2):59–81, 2014. 1

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

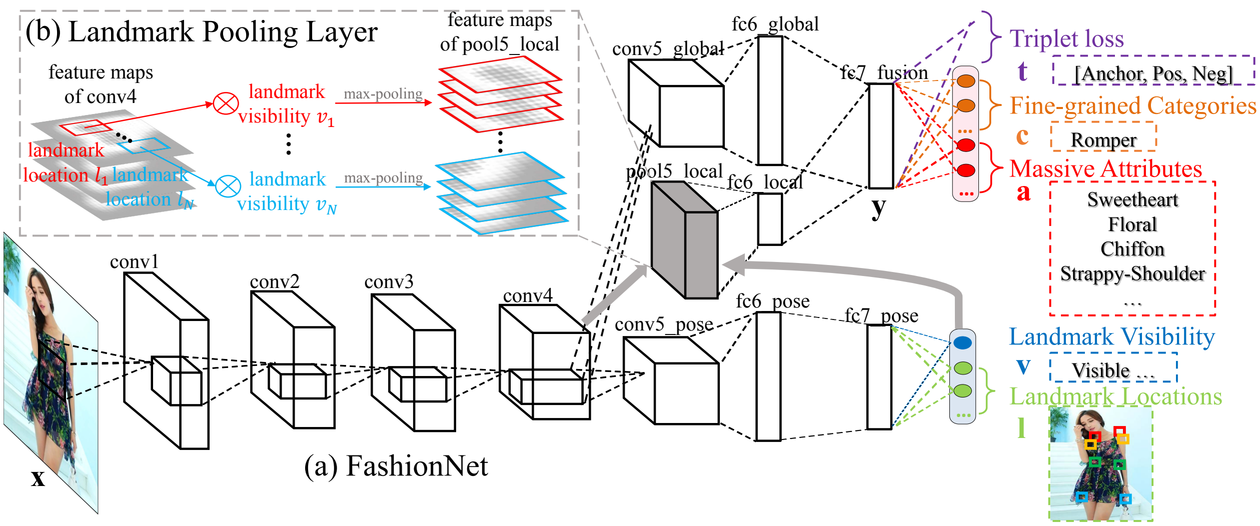

Figure 1: The detailed pipeline of FashionNet.

conv1	*conv2*		*conv3*		*conv4*		*conv5_pose*			*fc6&7_pose*	*loc.*	*vis.*
2×**conv**	**pool**	2×**conv**	**pool**	3×**conv**	**pool**	3×**conv**	**pool**	3×**conv**	**pool**	2×**fc**	**fc**	4×**fc**
3-1	2-2	3-1	2-2	3-1	2-2	3-1	2-2	3-1	2-2	-	-	-
64	64	128	128	256	256	512	512	512	512	1	1	1
relu	idn	relu	idn	relu	idn	relu	idn	relu	idn	relu	lin	soft
224	112	112	56	56	28	28	14	14	7	1024	8	2

conv5_global			*fc6_global*	*pool5_local*	*fc6_local*	*fc7_fusion*	*att.*	*cat.*
pool	3×**conv**	**pool**	**fc**	**lpool**	**fc**	**fc**	**fc**	**fc**
2-2	3-1	2-2	-	-	-	-	-	-
512	512	512	1	512×8	1	1	1	1
idn	relu	idn	relu	idn	relu	relu	sigm	soft
14	14	7	4096	4	1024	4096	1000	50

Table 4: The network architecture of FashionNet. Each table contains five rows, representing the '**name of layer**', '**receptive field of filter**'−'**stride**', '**number of output feature maps**', '**activation function**' and '**size of output feature maps**', respectively. Furthermore, '**conv**', '**pool**','**lpool**' and '**fc**' represent the convolution, standard max pooling, landmark max pooling and fully connection, respectively. Moreover, 'relu', 'idn', 'soft', 'sigm', and 'lin' represent the activation functions, including rectified linear unit [1], identity, softmax, sigmoid, and linear, respectively.

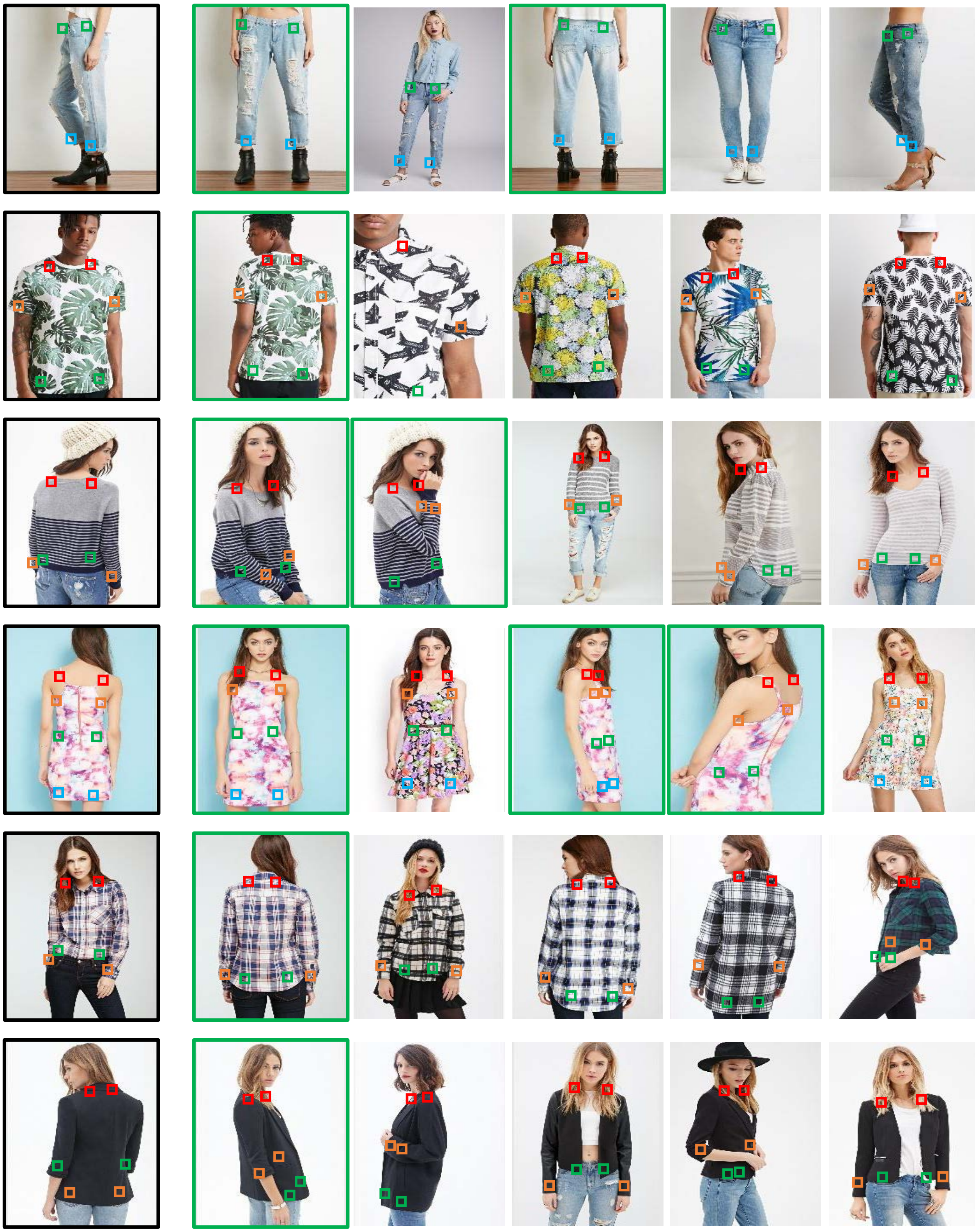

Figure 2: Visual results on in-shop clothes retrieval benchmark. Example queries, top-5 retrieved images, along with their predicted landmarks. Correct matches are marked in green.

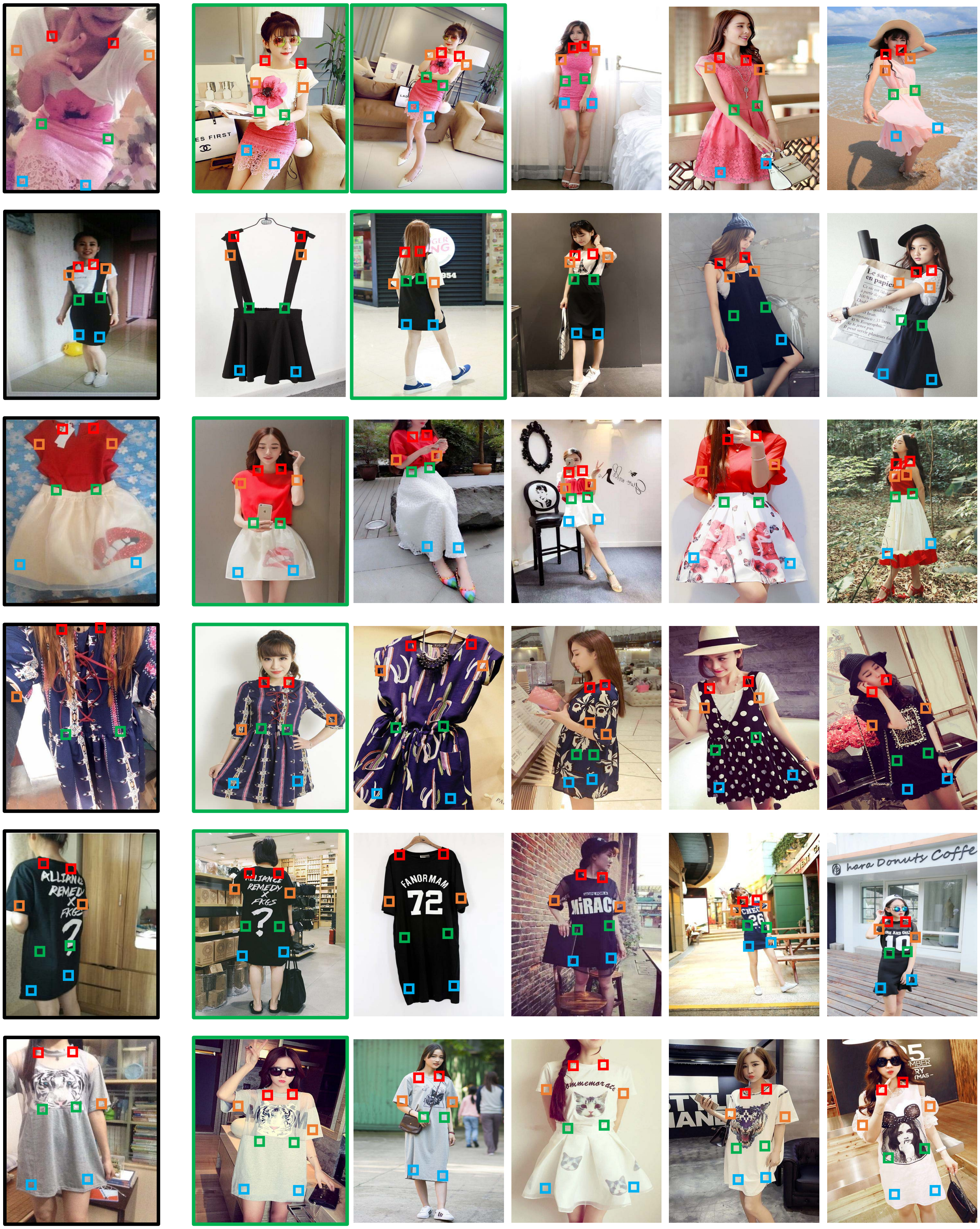

Figure 3: Visual results on consumer-to-shop clothes retrieval benchmark. Example queries, top-5 retrieved images, along with their predicted landmarks. Correct matches are marked in green.