A Dataset overview

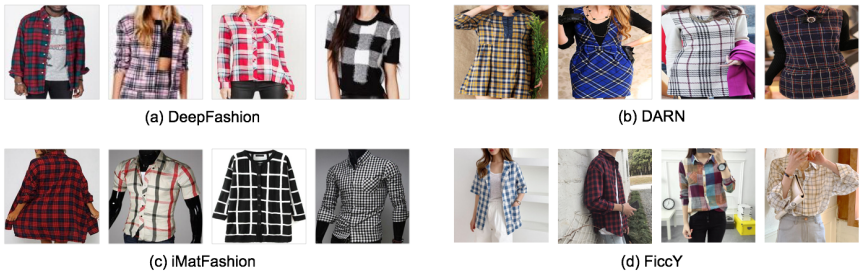

Fig. 1. Samples of ground-truth "check" in each benchmark datasets (DeepFashion, DARN, iMatFashion and FiccY). The fact the quality of the image (*e.g. view point, background noise etc.*) significantly differs according to the dataset highlights the importance of domain adaptability.

B Effect on training the student varying the amount of unlabeled data

Table 1. The comparison of mean AP (mAP) between the teacher (T) and the student (S) varying the amount of unlabeled images used for training.

mAP	T	S (Multi-task learning with 4 teachers)					
Train Size	-	1K	5K	10K	50K	100K	500K
Category	**85.50**	45.38	64.08	73.76	84.43	**86.04**	**87.18**
Pattern	**72.54**	13.11	41.46	57.80	70.72	**72.9**	**73.72**
Color	**58.58**	14.54	50.60	54.12	57.76	58.48	**59.33**
Texture	**62.30**	40.90	50.61	56.57	**62.97**	**63.75**	**64.20**

C Evaluation Metrics

The proposed model was mainly evaluated using the two most frequently used metrics: recall and F1 score. Recall is also referred to as sensitivity and it measures the probability of a positive detection. Precision, which is frequently used in combination with recall, is the percent of all relevant results among returned predictions. The F1 score is the harmonic mean of prediction and recall. The F1 score can fluctuate depending on which confidence score is used for a class to be considered as a final prediction. To minimize misleading effect by thresholding strategy, predictions were sorted by score and the top-k classes were selected as final predictions, meaning the number of predictions for an image was always k.