A detect-then-retrieve model for multi-domain fashion item retrieval

Michal Kucer
Rochester Institute of Techonology, USA*
mxk7721@rit.edu

Naila Murray
Naver Labs Europe, France
naila.murray@naverlabs.com

Abstract

Street-to-Shop fashion item retrieval is an instance-level image retrieval task in which a photo from a user is used to query a fashion image database in order to retrieve either the same or similar fashion items. This task is particularly challenging due to the domain shift between database photos, which tend to be staged professional shots, and consumer photos that have a much greater variety in terms of quality, pose, etc. To reduce the problem difficulty, state-of-the-art approaches train one retrieval model per domain or fashion item category. In this work we propose a single detect-then-retrieve model that can be applied to any (query or database) image and which outperforms methods using domain or category-specific retrieval models by significant margins on the Exact Street2Shop benchmark dataset.

1. Introduction

The past decades have seen the rise of e-commerce as a popular alternative to shopping in brick-and-mortar stores. More recently, applications such as NAVER shopping allow customers to search for items using images taken by their smartphone's camera. This task can be cast as an instance-level image retrieval task for the shopping domain [12, 19]. In this paper we address this task, and focus on fashion item retrieval in particular.

Fashion item retrieval using images provided by consumers as queries is particularly challenging due to the significant domain gap between these photos and photos taken by retailers. This domain gap arises because photos from retailers tend to be of much higher quality, in terms of lighting, resolution, and visual simplicity (e.g. with respect to clutter and occlusions). An illustration of this gap can be shown in Figure 2.

Another challenge is that fashion items such as clothing are highly deformable, such that their appearance exhibits high intra-instance variation. Due to these challenges, it is typical to improve the accuracy of methods for Street-to-Shop image retrieval by training domain-specific models [12] or training one model per fashion item category [19]. One assumes that the category of each database and query item is known and, for a given query image, the appropriate domain-specific image retrieval model for that category is used. State-of-the-art retrieval models are trained to generate representations for images that, when compared using a simple metric such as the cosine similarity, reflect the similarity of the image content [8]. The image database is therefore stored as a set of image representations extracted from the trained model. Using one model per category requires to store one representation per category, which is not desirable. In this work, we propose a cross-domain image retrieval model (Fig. 1) which outperforms per-category or domain-specific models while using one model for all categories and for both query and database images. Our model uses a detection model for fine-grained clothing item detection to reduce ambiguity in the retrieval objective. We train our retrieval model using both the standard triplet loss [8]

Figure 1. Schematic of the proposed detect-then-retrieve model. The framework consists of two steps: detection and retrieval. The query image along with the desired clothing category is passed into a clothing detection model to generate clothing detection proposals (a). The bounding box whose category matches the desired category is selected (b) and then passed into the retrieval model, which computes the image embedding used for shop image retrieval (c).

*Work conducted during an internship at Naver Labs Europe.

and the recently-introduced average precision (AP) loss [1] and find that models trained using these losses are complementary and can be effectively ensembled to boost performance. We make the following contributions:

- We introduce a detect-then-retrieve model which first uses a state-of-the-art object detection model trained to detect fashion items.
- We show that ensembling models trained using two complementary losses boost performance.
- We demonstrate that our single detect-then-retrieve model outperforms per-category model baselines by significant margins.

In section 2, we discuss related work on cross-domain fashion image retrieval. In sections 3 & 4, we describe our proposed detect-then-retrieve method and evaluate our method and provide quantitative and qualitative results.

2. Related work

As our work can be considered an application of instance-level image retrieval we first discuss general image retrieval methods before focusing on works related to cross-modal fashion image retrieval.

Image Retrieval. Traditional approaches to image retrieval typically adopt the following procedure: (i) extract local image features descriptors (e.g. scale-invariant feature transform (SIFT) [18]); (ii) embed them into a high-dimensional space using encoding techniques such as Bag-of-Visual-Words(BoVW) [3], or Fisher Vectors (FV) [22]; (iii) aggregate them to produce a fixed-length global representation; and (iv) apply a (perhaps learned) similar metric between representations to measure relevance [23]. More recently, convolutional image representations have achieved state-of-the-art results in image retrieval [8]. [25] showed the suitability of off-the-shelf features for image retrieval. It was shown that classification fine-tuning can further improve the quality of CNN features [2]. Current state-of-the-art approaches formulate image retrieval as a ranking problem and use an appropriate loss to optimize the order in which the images appear with respect to a query. Typical works in this vein use two- or three-stream Siamese network architectures combined with pairwise [20], triplet or n-tuplet losses [8] to train representations in an end-to-end manner. However, such approaches often require setting appropriate margins and using sophisticated hard-negative mining techniques [8, 29]). Recently, alternative loss functions based on optimizing for evaluation metrics such as average precision (AP) [1, 10] have been proposed and shown to achieve state of the art results on retrieval tasks, including instance-level image retrieval. In our work, we investigate the use of both the triplet loss and the AP loss and find that they are complimentary.

Cross-modal fashion image retrieval. Early exploration of cross-domain image retrieval [5, 6, 17] focused on retrieving "similar" clothing by collecting sets of images labeled with various clothing-related attributes. However, while such works have defined the similarity between two images as the number of high-level attributes in common, defining clothing similarity more rigorously is challenging. Kiapour et al. [19] and Huang et al. [12] were among the first to introduce the problem of exact street-to-shop clothing image retrieval, where the goal is to retrieve "shop" images that contain the exact item shown in the query image. Kiapour et al. [19] use a multi-layer perceptron (MLP) to learn a similarity measure between the "street" and "shop" image descriptors by minimizing the cross-entropy loss over pairs of CNN features which consist of street and shop images with matching or non-matching product IDs. Wang et al. [27] use a Siamese network architecture optimized simultaneously with a robust contrastive loss for image retrieval, and cross-entropy loss to regularize the network by predicting the 21,841 fine-grained categories of the images in the ImageNet dataset [4].

The most popular approach to optimizing image representations for retrieval uses the triplet loss. Huang et al. [12] proposed a dual attribute-aware ranking network (DARN), consisting of two networks each adapted to its specific image domain - "street" user images or shop images. These networks are used both for predicting semantic image attributes and image retrieval. Each image is first pre-cropped with a foreground clothing detector, after which the images are fed into the DARN network, which uses both the softmax loss to optimize attribute prediction and triplet loss to align the representations of images containing the same product. Liu et al. [33] propose the FashionNet model which jointly optimizes objectives for landmark prediction, category and attribute classification, and retrieval. The model first predicts the clothing landmarks, which are then used to pool and/or gate local features over estimated clothing landmarks. The local features are then concatenated with features from the whole clothing image for joint prediction of categories, attributes and retrieval (learned by optimizing the triplet loss). Jiang et al. [14] propose a bi-directional cross-triplet embedding for the task of cross-domain retrieval. More specifically, they break down the triplet loss and assign different weights to intra- and cross-domain losses. The network is fine-tuned for each category separately, with the convolutional layers being frozen, and only the last three fully-connected layers being fine-tuned with the proposed loss. Ji et al. [13] propose a network architecture which uses an attention mechanism to bias the pooling across the spatial regions, with different sub-networks for street and shop images. Gajic et al. [7]

Figure 2. Examples of the images contained in the *Exact Street2Shop* dataset, demonstrating the domain shift between user queries (left) and database images (right).

train a three column Siamese network using the triplet loss, in which they separate the streams according to whether they belong to the street set or the shop set, adapting the weights for each domain individually.

Previous works in cross-domain fashion retrieval have explored various ways to remove background clutter and focus on the main subject of the image. Kiapour et al [19] use selective search to generate high-confidence region proposals in their attempt to remove background clutter. Huang et al. [12] use selective search and an R-CNN model to crop clothing from images using humans as cues, but without considering clothing categories. Liu et al. [33] explore variations of the FashioNet model in which they compare using fashion landmark regression, human joint detection or body part detection to gate and/or pool features from an image. Zhang et al. [31] describe a weakly-supervised joint detection and retrieval system for image retrieval that considers various categories, e.g. dresses, furniture and toys. Different from these works, our framework trains a detector for fine-grained clothing detection, as opposed to using a global clothing detector [12], and uses it to select a single region to represent an image. Additionally, we consider the retrieval of multiple clothing categories using a single model, and do not limit ourselves to images of upper-body clothing [12].

3. Our approach

Our approach aims to tackle the scenario described in [19], in which the following assumptions are made:

- the ground-truth fashion item categories are known for both query and database images;
- the ground-truth bounding box for the query image is provided by a motivated user of the service;
- the ground-truth bounding box for database images are unknown.

We propose an efficient detect-then-retrieve approach that consists of a two-stage pipeline: (i) clothing item detection and (ii) clothing item retrieval. We describe each of these stages next.

3.1. Clothing item detection

We train a clothing item detector using the Mask R-CNN [9] detection architecture. This detector is trained to detect different clothing categories, as opposed to detecting one generic "clothing" category. Once trained, we apply the detector to all database images. To obtain a single crop for each image, it is first fed into the detector to produce a set of detection proposals. Each proposal has an objectness score and a predicted category. We filter the proposals to keep only those proposals with an objectness score greater than 0.5 and that have been categorized as belonging to the ground-truth category for that image. We then select, from this filtered list, the proposal with the highest category score. If the filtering process produces no proposals (*i.e.* if there are no proposals with an objectness score greater than 0.5 that have been predicted as belonging the ground-truth category for that image) then the bounding box is set to be the whole image. The resultant bounding box is then used to crop the database image.

3.2. Clothing item retrieval

Network architecture. Our baseline retrieval architecture is based on the end-to-end RMAC pipeline [8]. We used ResNet50 [11] as our baseline feature extractor f_θ, which generates a feature map $\mathbf{X} = [x_1; \ldots; x_k]$, where $x_i \in \mathbf{R}^N$ is the feature descriptor of the image I at the spatial location i. To aggregate the various descriptors, we replace R-MAC pooling with Generalized Mean (GeM) pooling [24],

$$\mathbf{x} = \left(\frac{1}{k} \sum_{i=1}^{k} (x_i)^p \right)^{\frac{1}{p}} \quad (1)$$

to obtain a single descriptor for each image. The descriptor is then fed through an l_2 normalization layer, a fully connected layer, and another l_2 normalization layer, to produce the final embedding $\mathbf{x}_I = f_\theta(I)$. I is the query or the database image, f_θ is a parametric function that computes the image embedding, and θ are the trainable parameters of f. Figure 3 shows the schematic of the approaches used to optimize the weights using the triplet and AP loss functions, which we describe next.

Learning with the triplet loss

We train a first model using the three-stream Siamese network architecture, which accepts a triplet of images: a query, a positive example (in our case another image with a

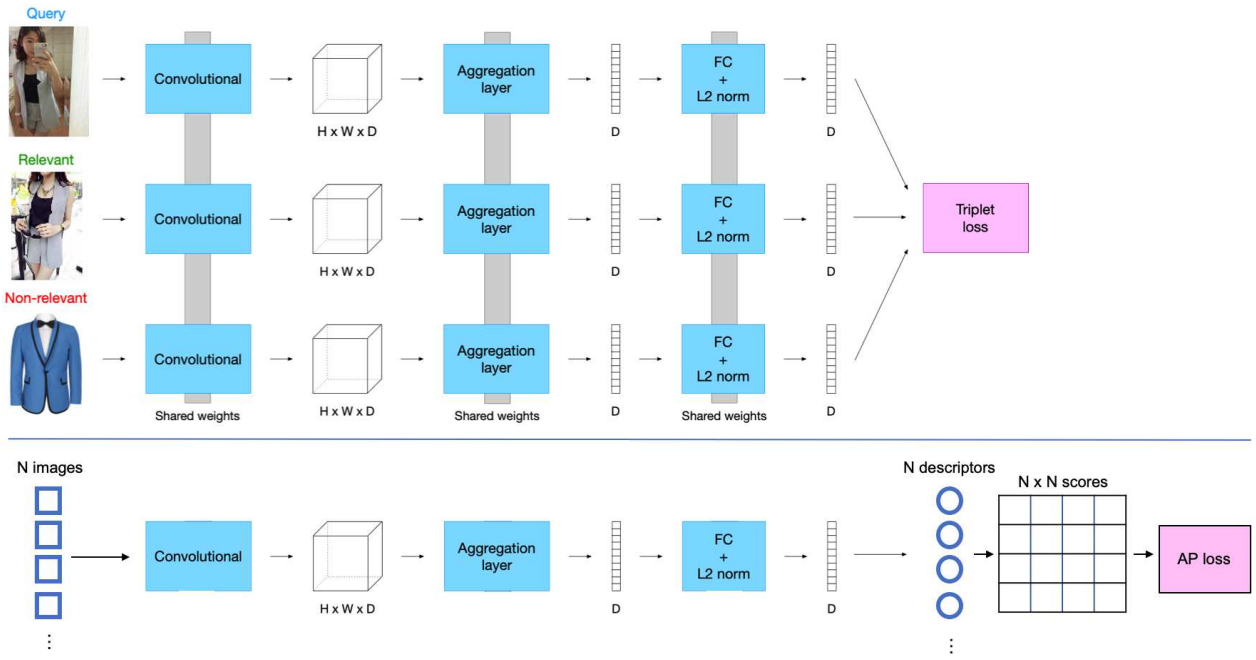

Figure 3. Schematic of architectures for proposed retrieval model trained with a triplet loss (top) or AP loss (bottom).

matching product ID), and a negative example. The weights between the different columns are shared.

Gordo et al. [8] show the improvement in ranking performance when the network is pre-trained with a classification loss. Therefore the network is first optimized to predict the product ID from an image. Afterwards, the weights are optimized with the triplet loss,

$$L(I_q, I^+, I^-) = \frac{1}{2} \max\left(0, m + \left\|q - d^+\right\|^2 - \left\|q - d^-\right\|^2\right), \tag{2}$$

where I_q and q, I^+ and d^+, and I^- and d^- are the image and feature descriptors for the query, positive, and negative images, respectively. This loss encourages the following property to hold: $sim(q, d^+) < sim(q, d^-) - \rho$ [10].

Learning with the AP loss

We also consider the AP loss described in [1] which directly optimizes the average precision (AP) for each query example. Let $\mathcal{X}$ be the set of all image representations and let $\mathcal{Q} \subset \mathcal{X}$ and $\mathcal{S} \subset \mathcal{X}$ be the sets of query and database representations respectively. Given a user "street" query $q \in \mathcal{Q}$, let S_q^+ and S_q^- be the sets of database images with matching and non-matching product IDs respectively. Given a list of items $r_i \in S_q^+ \cup S_q^-$ sorted by their increasing distance to q, average precision (AP) is defined as:

$$Prec@K = \frac{1}{K} \sum_{i=1}^{K} 1\left[r_i \in S_q^+\right] \tag{3}$$

$$AP = \frac{1}{\left\|S_q^+\right\|} \sum_{K=1}^{N} 1\left[r_k \in S_q^+\right] \cdot Prec@K \tag{4}$$

as defined in [10]. Though the AP metric cannot be directly optimized as it is non-differentiable, one can use the histogram binning approximation introduced by Ustinova and Lempitsky [26] to obtain a differentiable loss.

Model ensembles

Creating ensembles of deep models is known to improve performance significantly for tasks such as image classification [15]. To generate an ensemble image representation r for image retrieval, we perform feature-level fusion using representations r_{tl} and r_{apl} extracted from two models, one trained using the triplet loss and the other trained using the AP loss. To obtain a single representation we perform feature-level fusion by concatenating both as $\tilde{r} = [r_{tl}; r_{apl}]$. We then l_2-normalize $\tilde{r}$ to obtain r. In section 4 we compare the performance of both individual representations and their ensemble.

3.3. Querying the database

Each database image is cropped as described in section 4.2 and an image representation d is extracted using the model described in 3.2. In section 4 we evaluate the impact of applying the detection module to the database on retrieval performance.

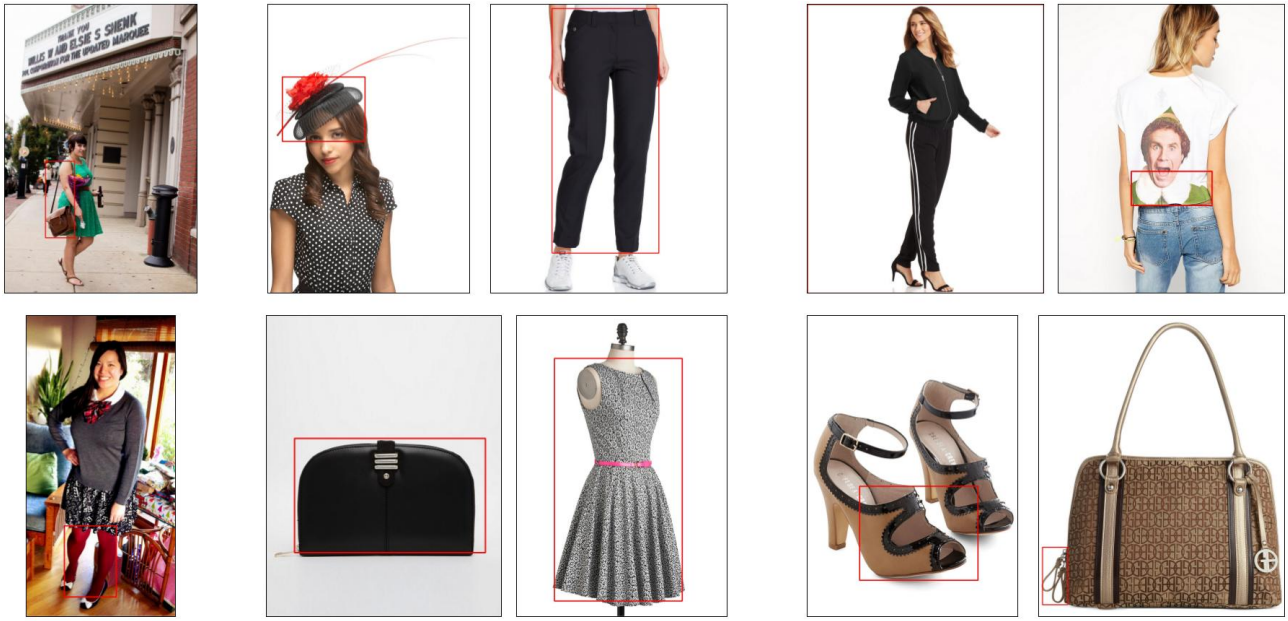

Figure 4. Detection examples for the query (left) and database (correct (middle) and incorrect(right)) images in the Street2Shop dataset.

To conduct a query, the query image is cropped using its ground-truth bounding box and its representation q is extracted using the retrieval model. The similarity between query descriptor q and database descriptor d_i is computed as the inner product between their embeddings:

$$sim(q, d_i) = q^\top d_i. \tag{5}$$

These similarity scores can then be sorted to return a list of decreasingly relevant database items.

4. Experiments

4.1. Datasets

ModaNet [32]. ModaNet is a recently introduced large-scale dataset that contains street images of fashion items. The dataset consists of a total of 55,176 images and provides polygon labels for 13 types of clothing items. The dataset currently contains annotations only for the training split of the dataset. Therefore we randomly chose 5% of the images to use as a validation set. Note that this dataset does not contain product ID information and thus cannot be used to train and/or evaluate retrieval models. We use it to train our detector.

Exact Street2Shop. Kiapour et al. (2015) introduced the *Exact Street2Shop* [19] dataset to tackle the challenging problem of Street-to-Shop clothing retrieval. Though various datasets, like DARN [12], were released for solving this problem, we chose the *Exact Street2Shop* dataset as it contains over 400,000 shop images, over 20,000 "street" images, and images have been labelled with 11 different clothing categories. In addition, the query images have been annotated with bounding boxes of clothing items. Figure 2 shows examples of "street" and "shop" images from the *Exact Street2Shop*.

4.2. Clothing item detection

Though Exact Street2Shop dataset contains bounding boxes for the query items, there is only roughly 40,000 of boxes as compared to the ModaNet, which contains a 55,000 images each labeled with various piece of clothing. Thus we trained our Mask R-CNN [9] clothing item detector using the ModaNet dataset. In order to ensure the quality of the detector, we use the AP at 50% Intersection over Union (IoU) metric to evaluate its performance. Though the validation sets are not equivalent and results not directly comparable, the detector achieves similar overall and per-category quantitative performance as compared to the best detection models reported in [32]. In particular, our model achieves an overall mean AP of 0.893, as compared 0.82 mAP [32]. We used the Mask R-CNN model with a Feature Pyramid Network (FPN)[16] backbone based on the ResNext-101 architecture.

Qualitative Results In Figure 4, we show detection examples for the "street" and "shop" domains from the *Exact Street2Shop* dataset which is used for retrieval training and experiments. As is expected, fashion item detection for query images is quite good as they are similar to images in the ModaNet dataset. For the case of "shop" images we experience a slight domain shift as these images are a mixture

	metric	Tri DB F	Tri DB C	mAP DB F	mAP DB C	Tris	mAPs	Fulls	Crops
bags	mAP	0.1819	0.1928	**0.2262**	0.1914	0.2082	0.2238	**0.2684**	0.2339
	Acc@1	0.2518	0.3094	**0.3309**	0.2806	0.3022	0.3237	**0.3813**	0.3597
	Acc@20	**0.6115**	0.5612	**0.6115**	0.5827	0.6043	**0.6691**	0.6619	0.6259
belts	mAP	0.0674	0.0809	0.0778	**0.0874**	0.1024	**0.1072**	0.1005	0.0942
	Acc@1	0.0714	0.0952	0.0476	**0.1190**	**0.1429**	**0.1429**	0.1190	0.0952
	Acc@20	0.3571	0.3095	0.2857	**0.3810**	0.3810	0.3333	0.3571	**0.4286**
dresses	mAP	0.4592	**0.4644**	0.4306	0.4253	0.4866	0.4548	**0.5091**	0.4954
	Acc@1	0.5166	**0.5208**	0.4834	0.4984	0.5403	0.5176	0.5606	**0.5640**
	Acc@20	**0.7162**	0.7159	0.6674	0.6593	0.7302	0.6806	**0.7407**	0.7204
eyewear	mAP	0.2121	0.2084	0.1167	**0.2371**	0.2189	0.2053	0.1723	**0.2671**
	Acc@1	0.2586	0.2241	0.1379	**0.3103**	0.2586	0.2586	0.2069	**0.3621**
	Acc@20	0.7069	0.8621	0.5862	**0.8793**	0.7759	0.6724	0.6897	**0.9138**
footwear	mAP	**0.1301**	0.1048	0.0719	0.0749	0.1294	0.0803	**0.1382**	0.1103
	Acc@1	**0.1539**	0.1384	0.0819	0.1052	**0.1600**	0.0963	0.1573	0.1478
	Acc@20	**0.4048**	0.3671	0.2204	0.2453	**0.4103**	0.2464	0.3771	0.3416
hats	mAP	0.2109	**0.2957**	0.1660	0.2761	0.2765	0.2574	0.2219	**0.3274**
	Acc@1	0.2000	**0.2615**	0.2000	0.2462	0.2462	0.2308	0.2462	**0.3077**
	Acc@20	**0.7077**	**0.7077**	0.5077	0.6769	0.7077	0.6769	**0.7231**	0.7077
leggings	mAP	0.1330	0.1563	0.1573	**0.1606**	0.1510	0.1705	**0.1875**	0.1818
	Acc@1	0.1510	0.1766	0.1852	**0.1937**	0.1681	0.1994	**0.2251**	0.2051
	Acc@20	**0.4758**	0.4729	0.3533	0.3789	**0.5071**	0.3704	0.4843	0.4900
outerwear	mAP	0.2088	0.2241	**0.2645**	0.2497	0.2323	0.2739	0.2771	**0.2806**
	Acc@1	0.2348	0.2530	**0.3018**	0.2835	0.2530	**0.3079**	0.2988	0.3049
	Acc@20	0.4329	0.4482	**0.4726**	0.4451	0.4634	0.4695	**0.4909**	0.4787
pants	mAP	0.2081	**0.2626**	0.2251	0.2484	0.2341	0.2580	0.2310	**0.2817**
	Acc@1	0.2727	**0.3333**	0.2727	0.2727	0.2879	0.3030	0.2727	**0.3333**
	Acc@20	0.4091	**0.5000**	0.3788	0.4848	0.4545	0.4242	0.4091	**0.5152**
skirts	mAP	0.5420	**0.5862**	0.5433	0.5829	0.5968	0.5836	0.5890	**0.6231**
	Acc@1	0.6091	**0.6320**	0.5711	0.6294	0.6294	0.5990	0.6142	**0.6802**
	Acc@20	0.7919	**0.7944**	0.7716	0.7741	0.7970	0.7944	**0.8173**	0.8020
tops	mAP	0.3187	0.2988	**0.3488**	0.3265	0.3374	0.3600	**0.4060**	0.3687
	Acc@1	0.3639	0.3425	**0.3945**	0.3777	0.4021	0.4128	**0.4480**	0.4266
	Acc@20	0.5734	0.5933	**0.6147**	0.5765	0.6101	0.6009	**0.6407**	0.6162
average	mAP	0.2429	**0.2614**	0.2389	0.2600	0.2703	0.2704	0.2819	**0.2967**
	Acc@1	0.2804	0.2988	0.2734	**0.3015**	0.3082	0.3084	0.3209	**0.3442**
	Acc@20	0.5625	**0.5757**	0.4972	0.5531	0.5856	0.5398	0.5811	**0.6036**

Table 1. Comparison of the mAP, Top-1, and Top-20 retrieval accuracy for the *Exact Street2Shop* dataset. The first four models represent the ablation experiments in which the loss functions and database images were varied. Tri / mAP indicated whether a triplet loss or AP loss was used in training of the network. DB F / C indicates whether the database images were un-cropped or cropped respectively. The last four columns report the results for the various ensemble representations considered in our experiments. We list each ensemble and models it combines in parentheses: *Tris* (*Tri DB F* / *Tri DB C*), *mAPs* (*mAP DB F* / *mAP DB C*), *Fulls* (*Tri DB F* / *mAP DB F*), and *Crops* (*Tri DB C* / *mAP DB C*). Best models are highlighted in bold for both single (first four columns) and ensemble models.

of images with people modelling the item and images that display the item by itself on a uniform background.

4.3. Image Retrieval

In our retrieval experiments, we explore the following aspects: (a) the effect of using the fine-grained clothing item detector to crop database images on the retrieval accuracy; (b) a comparison of the triplet and AP losses; and (c) the effect of combining learned representations on the retrieval performance.

We Followed the evaluation protocol of Kiapour et al. [19], the experiments are restricted to within category retrieval. For each set of experiments, we report the mean average precision (mAP), defined by Eq. 4 and the Top-k retrieval accuracy ($Acc@K$):

$$Acc@K = \frac{1}{N}\sum_{i=1}^{N} 1\left[\mathcal{S}_q^+ \cap \mathcal{S}_q^K\right], \tag{6}$$

where $1\left[\mathcal{S}_q^+ \cap \mathcal{S}_q^K\right]$ is an indicator function that equals 1 if the set of the top-K retrieval images contains a database image that matches the product ID of the query image.

Implementation details

The retrieval models and both losses are implemented in Python using the PyTorch [21] framework. During training,

each image either has its smallest side (AP loss) or largest side (triplet loss) re-sized to 800 pixels and is augmented with the following set of image transformations: color distortion, random tilting, random skew, and random cropping to 800×800 (AP loss). Each model was initialized with the network weights pre-trained on the ImageNet dataset and trained until convergence. For the triplet loss, we follow the weight update scheme of Gordo et al. [8], which allows for use of high-resolution images in training of the network. Given an image triplet, the gradients of L with respect to d, d^+, and d^- are computed sequentially and aggregated over the triplet and the batch of size b. For the AP loss, we follow the weight update scheme outlined in [1], allowing for use of large images and arbitrary batch sizes. For both the triplet and AP losses, images from arbitrary categories are used in mini-batches. That is, we do not train on a per-category basis but ignore category information during the retrieval training phase and train a single model. This model is then applied on a per-category basis during the testing phase.

Effect of object localization

In the first set of ablation experiments, we explore the effect of obtaining the database crops on the learned representations. If you compare column 1 to column 2 and column 3 to column 4 of Table 1, we can see that on average cropping the database images aids the retrieval when trained with both the triplet loss and the AP loss. The performance gain is found to be more significant for the AP loss (columns 3 and 4). In particular we see significant improvements in retrieval performance for the eye-wear and pants categories, which are often modeled by a person or with a mannequin. Despite the good detector performance in the footwear category, the triplet loss loses some performance. The AP loss improves by 2 % in retrieval accuracy (footwear category results of columns 3 and 4), which is still worse than the triplet loss. This can be explained by examining the database images for the footwear category and seeing that most of the images are already very clean photos displaying the particular shoe product on a simple background, and thus there is not much to be gained by further cropping the image. Similarly, we see little to no improvement in retrieval accuracy in categories that often occupy a large portion of the image, such as dresses, skirts, and tops.

Effect of different retrieval losses

In the second set of ablation experiments, we explore the effect of different training loss functions on the learning process. Table 1 shows that training the retrieval network with the AP loss achieves similar results in terms of mAP and Top-1 retrieval accuracy, however we can see that networks trained with the AP loss show worse results in terms of Top-20 retrieval accuracy. We hypothesize that this is because the AP loss has a small gradient for misranked images later in the ranking (as such images have a small impact on the AP) and therefore the supervisory signal to the model during training is weak. This is perhaps mitigated by using detection, as we see that the drop in Top-20 accuracy from triplet to AP loss is much lower when the database images are cropped.

Categories	Single model			Per-category
	Wang et al. [27]	Our best model	Our best ensemble	Kiapour et al. [19]
Bags	46.6	56.1	**62.6**	37.4
Belts	20.2	31.0	**42.9**	13.5
Dresses	56.9	71.6	**72.0**	37.1
Eyewear	13.8	86.2	**91.4**	35.5
Footwear	13.1	**36.7**	34.2	9.6
Hats	24.4	**70.8**	**70.8**	38.4
Leggings	15.9	47.3	**49.0**	22.1
Outerwear	20.3	44.8	**47.9**	21.0
Pants	22.3	50.0	**51.5**	29.2
Skirts	50.8	79.4	**80.2**	54.6
Tops	48.0	59.3	**61.6**	38.1
Average	30.2	57.6	**60.4**	30.6

Table 2. Comparison of our best performing single and ensemble models to the state-of-the-art on the Street2Shop dataset.

Categories	VAM [28]	P. Non-Sh.[30]	P. Sh. [30]	Ours
Dresses	0.621	0.571	0.583	**0.716**
Outerwear	N/A	0.500	**0.509**	0.448
Skirts	0.709	0.736	0.723	**0.794**
Tops	0.523	0.467	0.470	**0.593**

Table 3. Comparison of the Top-20 retrieval accuracy for methods which considered a subset of the categories.

Effect of model ensembling

Table 1 shows results for different ensembles. As is expected, we observe noticeable improvements in mAP, Top-1 and Top-20 accuracy in all cases. Significant improvement can be obtained when two models trained with different losses are combined. Ensembles of models trained with either the triplet or AP loss, with either cropped and un-cropped images, improve the performance over their respective single models with cropped database images. A more significant improvement is achieved when one of the models is trained with the triplet and the other with the AP loss. When combining models trained using the AP loss and triplet loss (and each trained with cropped images), we see absolute improvements of 3.5 % mAP, 4.5 % Top-1 accuracy, and 2.8 % Top-20 accuracy when compared with the best individual model trained on cropped images.

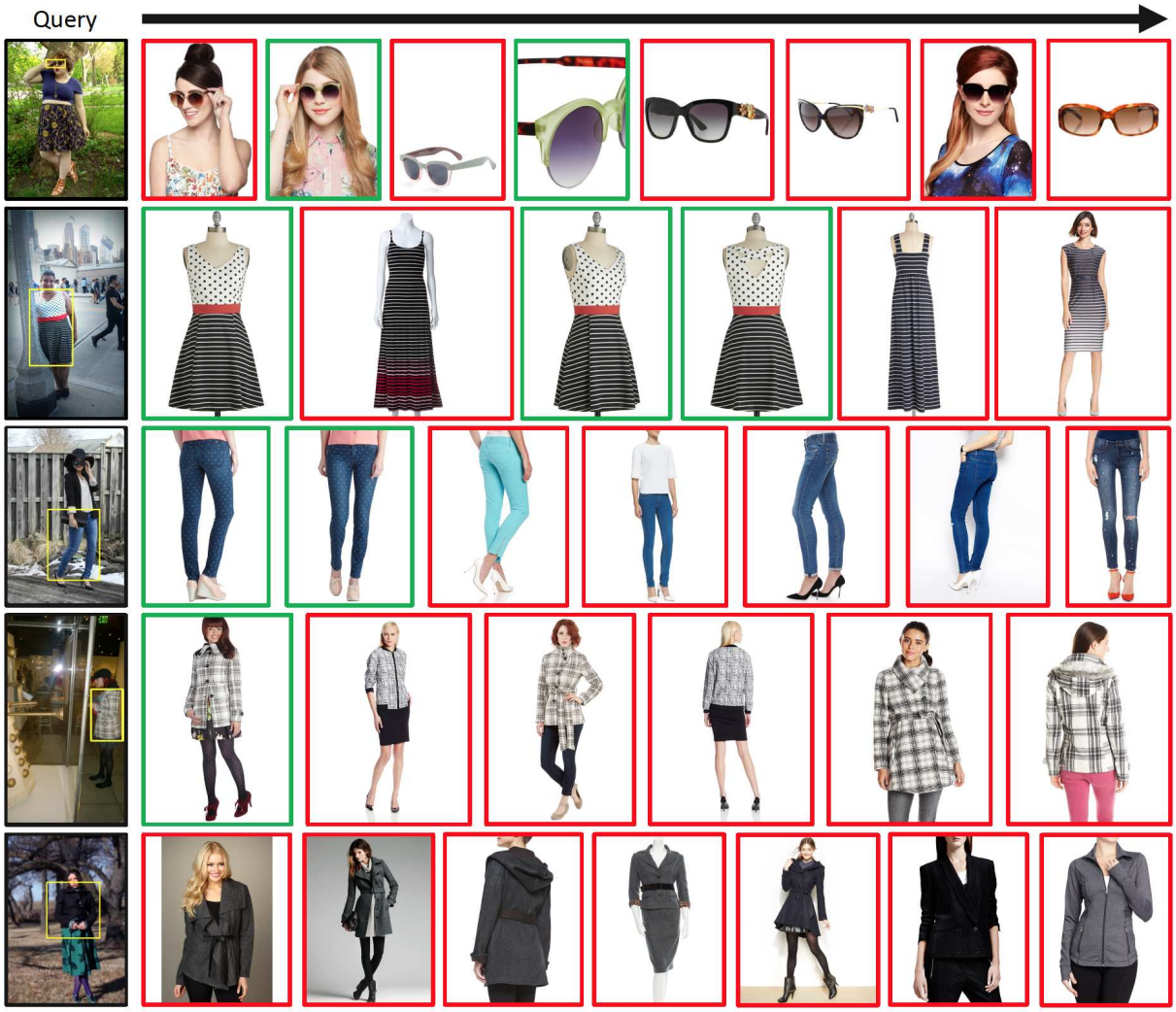

Figure 5. Qualitative retrieval examples. The images in the left column are examples of image queries with their ground-truth bounding boxes shown in yellow. To the right of them are database images ordered by decreasing similarity to the query. Images with *green* borders match the product in the image. The last row shows an example of an image for which there was no ground truth match in the top 7 retrieved images.

Comparison to the state-of-the-art

Table 2 compares the best single and ensemble models to the previous state-of-the approaches that either (a) trained a single model for all categories, or (b) used a set of models that were fine-tuned per category. Both of our models perform significantly better in terms of Top-20 retrieval accuracy per-category as well as overall.

Table 3 compares our best single model with works that only trained and evaluated their models on a subset of the categories of the *Exact Street2Shop* dataset. Our best single model outperforms the results of [28] and [30] for the dresses, skirts, and tops categories. Note that [28] and [30] train separate models for each category.

5. Conclusion

In this paper, we propose a memory-efficient detect-then-retrieve framework for cross-modal fashion image retrieval, which consists of fine-grained clothing detection followed by retrieval. We show that our framework achieves state-of-the-art results and outperforms category-specific models. Additionally, we explored the retrieval performance of our models and showed that the triplet and AP loss are complementary and, when combined, lead to significant performance gains.

References

[1] Anonymous. Learning with average precision: Training image retrieval with a listwise loss. *Under anonymous review*, 2019. 2, 4, 7

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. *CoRR*, abs/1404.1777, 2014. 2

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

[5] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–13, June 2013. 2

[6] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part II*, ACCV'12, pages 420–431, Berlin, Heidelberg, 2013. Springer-Verlag. 2

[7] B. Gajic and R. Baldrich. Cross-domain fashion image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[8] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, Sep 2017. 1, 2, 3, 4, 7

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 3, 5

[10] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. *CoRR*, abs/1804.05312, 2018. 2, 4

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[12] J. Huang, R. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1062–1070, Washington, DC, USA, 2015. IEEE Computer Society. 1, 2, 3, 5

[13] X. Ji, W. Wang, M. Zhang, and Y. Yang. Cross-domain image retrieval with attention modeling. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 1654–1662, New York, NY, USA, 2017. ACM. 2

[14] S. Jiang, Y. Wu, and Y. Fu. Deep bi-directional cross-triplet embedding for cross-domain clothing retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 52–56, New York, NY, USA, 2016. ACM. 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 4

[16] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 5

[17] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3330–3337, June 2012. 2

[18] D. Lowe. Object recognition from local scale-invariant features. volume 2, pages 1150 – 1157 vol.2, 02 1999. 2

[19] S. L. A. C. B. T. L. B. M. Hadi Kiapour, Xufeng Han. Where to buy it:matching street clothing photos in online shops. In *International Conference on Computer Vision*, 2015. 1, 2, 3, 5, 6, 7

[20] I. Melekhov, J. Kannala, and E. Rahtu. Siamese network features for image matching. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 378–383, Dec 2016. 2

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 6

[22] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 2

[23] F. Perronnin, Y. Liu, J. Snchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, June 2010. 2

[24] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *TPAMI*, 2018. 3

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014. 2

[26] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4170–4178. Curran Associates, Inc., 2016. 4

[27] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y.-G. Jiang. Matching user photos to online products with robust deep features. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 7–14, New York, NY, USA, 2016. ACM. 2, 7

[28] Z. Wang, Y. Gu, Y. Zhang, J. Zhou, and X. Gu. Clothing retrieval with visual attention model. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, Dec 2017. 7, 8

[29] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017. 2

[30] Y. Xiong, N. Liu, Z. Xu, and Y. Zhang. A parameter partial-sharing cnn architecture for cross-domain clothing retrieval. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4, Nov 2016. 7, 8

[31] Y. Zhang, P. Pan, Y. Zheng, K. Zhao, Y. Zhang, X. Ren, and R. Jin. Visual search at alibaba. In *Proceedings of the 24th*

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, pages 993–1001, New York, NY, USA, 2018. ACM. 3

[32] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *ACM Multimedia*, 2018. 5

[33] S. Q. X. W. Ziwei Liu, Ping Luo and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3