Unsupervised Meta-Domain Adaptation for Fashion Retrieval

Vivek Sharma[1,3,4], Naila Murray[2], Diane Larlus[2],
M. Saquib Sarfraz[1], Rainer Stiefelhagen[1], Gabriela Csurka[2]
[1] Karlsruhe Institute of Technology, [2] NAVER LABS Europe
[3] Massachusetts Institute of Technology, [4] Harvard Medical School

Abstract

Cross-domain fashion item retrieval naturally arises when unconstrained consumer images are used to query for fashion items in a collection of high-quality photographs provided by retailers. To perform this task, approaches typically leverage both consumer and shop domains from a given dataset to learn a domain invariant representation, allowing these images of different nature to be directly compared. When consumer images are not available beforehand, such training is impossible. In this paper, we focus on this challenging and yet practical scenario, and we propose instead to leverage representations learned for cross-domain retrieval from another source *dataset and to adapt them to the* target *dataset for this particular setting. More precisely, we bypass the lack of consumer images and directly target the more challenging meta-domain gap which occurs between consumer images and shop images, independently of their dataset. Assuming that datasets share some similar fashion items, we cluster their shop images and leverage the clusters to automatically generate pseudo-labels. Those are used to associate consumer and shop images across datasets, which in turn allows to learn meta-domain-invariant representations suitable for cross-domain retrieval in the target dataset. The features and code are available at* `https://github.com/vivoutlaw/UDMA`.

1. Introduction

Visual search is an increasingly popular functionality for querying large e-commerce catalogues, particularly for fashion. Fashion item retrieval differs from standard instance-level retrieval such as landmark retrieval [12, 38] and from person or vehicle re-identification [43, 30, 44]. In fashion retrieval, we need to deal with strong appearance variations: while consumer images used as queries can be of very low quality, shop images are in general aesthetically pleasing and iconic shots provided by retailers. Consequently there is a strong domain gap between consumer images and shop images. This task is known as consumer-to-shop (C2S) retrieval and is a challenging cross-domain retrieval task. Many approaches have been proposed to tackle this domain shift [21, 17, 51, 24]. These assume that both shop images and consumer images are available for training which is a realistic assumption when a retail e-store has existed for some time and examples of query-result pairs are available.

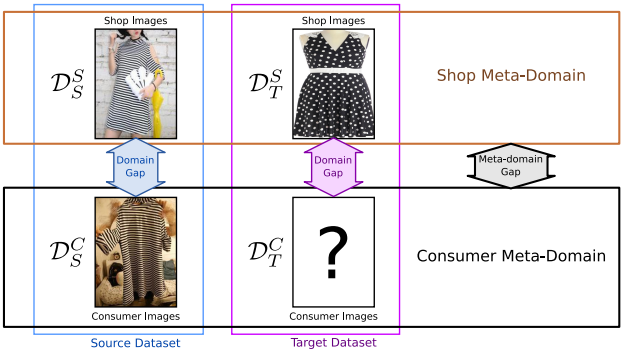

Figure 1. Cross-domain fashion item retrieval is tackled for a target dataset - for which consumer images are not available - by i) leveraging the cross-domain retrieval model of a source dataset and ii) tackling the meta-domain gap directly.

What happens when query-result pairs are not available? This arises naturally in practice, *e.g.* when a new shopping catalog is introduced, or when a store considers a new market in a different geographical region. When consumer images are not available beforehand, for the simple reason that user queries are not accessible before the service is deployed, traditional methods cannot be applied. In addition, collecting relevance labels from consumers can introduce privacy concerns, and collecting relevance labels using annotation campaigns is laborious and expensive.

In this work, we tackle this more challenging and realistic scenario by adapting a cross-domain retrieval model trained on a *source* C2S dataset to a *target* C2S setting, using what we call a *meta-domain adaptation* method. We define a **meta-domain** as a high-level grouping of similar sensor domains. For example, shop images from different sources all belong to a *shop meta-domain*. Similarly, the

consumer meta-domain consists of available consumer images from different data sources. In this setting, we assume access to at least one labeled (C2S) source set containing consumer and shop images paired according to their product ID (PID). We then propose to adapt the corresponding cross-domain C2S source model to a new unlabeled (C2S) target set for which training images are available only for a single domain (*e.g*. shop). We do this by approximating the domain gap between the shop and consumer images in this target set by the *gap between the meta-domains*. We can view this meta-domain gap as a generic gap between all the possible shop images and all the possible consumer images of a given product. In our case, it corresponds to differences in sensor modality and photographic style. Each domain, *e.g*. $\mathcal{D}_S^C$ or $\mathcal{D}_T^S$ (see Figure 1), is a *meta-domain instance*.

When we consider several datasets, additional mismatch might exist between domain instances within a meta-domain. This is primarily due to dataset sampling bias: different domain instances might span different subsets of fashion categories and products. Furthermore, there is no guarantee that the sampling is consistent across meta-domains, *i.e*. that the same categories and products exist in each dataset. Hence, a transformation that aligns the distributions of the shop domains (*e.g*. using classical unsupervised domain adaptation) cannot be directly used for aligning those of the consumer domains. Instead, we assume that there is an overlap between the shop domain distributions, and rely on similar product items to adapt the source model to the target by minimizing the meta-domain gap. To find overlapping items we jointly cluster the shop domains. By propagating PIDs from source shop items to target shop items in the same clusters, we can then pair consumer and shop images across meta-domains, irrespective of the datasets to which they belong. This unsupervised adaptation allows us to perform cross-domain retrieval when target consumer queries become available.

Our contribution is threefold: (1) We introduce the concept of meta-domain and the problem of meta-domain adaptation, and show how they are useful for approaching a challenging and realistic scenario of unsupervised cross-domain transfer, where we have access to only one of the two target domains (Section 3). (2) We propose an effective strategy to minimize the meta-domain gap by automatically finding class-aligned instances in the available labeled and unlabeled meta-domain instances (Section 4). (3) We experimentally show that this unsupervised meta-domain adaptation approach achieves good cross-modal retrieval results on the Street-to-Shop benchmark, without access to any consumer images from this dataset at train time (Section 5).

2. Related Work

Fashion Retrieval. For the easier shop-to-shop retrieval task, fashion retrieval is treated as standard instance-level retrieval. Initially based on aggregated local descriptors, recent approaches fine-tune CNNs for retrieval using ranking losses, *e.g*. contrastive [38], triplet [12] or AP loss [39].

The more challenging consumer-to-shop (C2S) retrieval task requires bridging the gap between the consumer and shop domains, and many techniques have been proposed. Most methods reduce this gap by using pairwise or triplet losses that mix images from both domains [21, 51, 17, 19, 18, 24]. One can further reduce the gap by learning street- and shop-specific image representations [17, 53, 18, 8]. Another way to reduce the domain gap is to explicitly remove the background of consumer images, which is one of the most critical sources of appearance variation. This approach is followed by [21], which uses object proposals to select foreground items, and [17, 24] who use object detectors for the same purpose. Lastly, recent methods have proposed pose estimation and part detection strategies to align and compare images at the part level. This can be achieved without supervision using attention mechanisms [18, 52] or graph reasoning [23], or with supervision by training and applying fashion landmarks, joint and/or body-part detectors [29, 32, 55]. None of these methods address the unsupervised meta-domain adaptation scenario.

Unsupervised Domain Adaptation. Early shallow domain adaptation (DA) methods include data reweighting, metric learning, subspace representations or distribution matching (see the surveys [11, 6]). The last few years have witnessed an increasing number of deep learning methods for DA. Many of them revisit earlier shallow methods using *e.g*. Siamese networks to minimize the discrepancy between feature distributions [33, 7, 47]. Adversarial methods include discriminative models with adversarial loss [48, 50] or gradient reversal layer [9]. Generative models are also frequently used to adapt models when the domain shift can be seen as an image style variation [28, 15]. Other methods directly adapt the network weights [40], batch normalization [4] or dropout regularization [41] layers. Closer to our work, [46, 54, 20, 5, 16] refine models using pseudo-labels generated in the target space. Some methods explicitly handle the case where the source and target domains only share a subset of their labels. Partial DA methods [2, 3, 25] tackle a scenario where unseen classes only belong to the source domain. Open set DA methods [35, 42, 27] are used when the target set contains unseen classes. They typically try to identify and exclude these classes from adaptation. In contrast, our approach deals exclusively with unseen classes (source and target domains share no product ID).

The above DA methods mainly focus on image categorization and only few works target instance-level or cross-domain retrieval. Amongst them, retrieval was mainly addressed in the context of person re-identification [34, 1, 56], the adaptation being between domains but the retrieval being performed within a single domain. [26] proposes cross-

Table 1. **Descriptions of different domain-adaptation (DA) and meta-domain adaptation (MDA) scenarios**. The standard consumer-to-shop (C2S) retrieval task only involves a target dataset. Standard DA could be considered to transfer a C2S model from a source to a target set if those sets shared product IDs. This is not the case for MDA. In this work, we consider the additional challenge that no consumer image is available for the target set. This constitutes the **Unsupervised MDA** scenario defined and tackled in this paper. We list an alternative unsupervised MDA (*) where the target shop image labels are available and can be used, not tackled in this paper.

Scenario name	Source dataset		Target dataset		Shared PIDs
	Images (Consumer,Shop)	Prod IDs (Consumer,Shop)	Images (Consumer,Shop)	Prod IDs (Consumer,Shop)	
	Consumer-to-shop domain adaptation in the target dataset				
No adaptation (NA)		-	(N,Y)	(N,Y)	-
Unsupervised DA (UDA)		-	(Y,Y)	(N,Y)	-
Supervised DA (SDA)		-	(Y,Y)	(Y,Y)	-
	Domain adaptation across consumer-to-shop retrieval datasets				
No adaptation (NA)	(Y,Y)	(Y,Y)	(N,N)	(N,N)	Y
Unsupervised DA (UDA)	(Y,Y)	(Y,Y)	(Y,Y)	(N,N)	Y
Supervised DA (SDA)	(Y,Y)	(Y,Y)	(Y,Y)	(Y,Y)	Y
	*The proposed **Meta-domain adaptation (MDA)** task*				
No adaptation (NA)	(Y,Y)	(Y,Y)	(N,N)	(N,N)	N
Unsupervised MDA	(Y,Y)	(Y,Y)	(N,Y)	(N,N)	N
Unsupervised MDA (*)	(Y,Y)	(Y,Y)	(N,Y)	(N,Y)	N
Supervised MDA	(Y,Y)	(Y,Y)	(Y,Y)	(Y,Y)	N

domain 3D model retrieval represented by multi-view 2D images. Most related to our work, the scene graph approach of [36] transfers knowledge from a source domain to improve cross-media retrieval in a target domain through media and distribution alignment. However, their approach assumes access to both modalities - images and text - from both datasets, and tackles DA within the same modalities.

3. Meta-domain Adaptation

This section presents the concept of meta-domain adaptation and precisely defines it in the context of our target application, *i.e*. consumer-to-shop (C2S) fashion item retrieval. It introduces the terminology used in the rest of the paper. It also discusses similarities and differences with other unsupervised and supervised domain adaptation tasks that could be defined in the context of C2S retrieval.

Context. In this work, our aim is to learn domain agnostic representations for the problem of consumer-to-shop (C2S) fashion item retrieval for a *target* dataset, by leveraging a C2S retrieval model learnt on a *source* dataset. We assume that a C2S dataset consists of a pair of sets: a set of consumer images and a set of shop images. For the particular scenario we are interested in, we assume that consumer images of the target dataset are unavailable during training.

Link with standard retrieval. C2S retrieval fundamentally differs from the classical retrieval task, for which the query and the retrieved images belong to the same domain. In our case, query images belong to the consumer domain while retrieved images belong to the shop domain.

Link with standard domain adaptation. C2S retrieval also differs from the classical domain adaptation task where a classifier trained on a dataset from the source domain is adapted to perform well on classification on the target domain. Here, our goal is to learn domain invariant representations for cross-domain retrieval instead.

3.1. Consumer-to-shop retrieval

Table 1 summarizes all the different scenarios which can arise when looking at the consumer-to-shop (CS) retrieval problem. For each scenario we precisely describe which data the training algorithm has access to. In what follows, we carefully describe connections with the DA and transfer learning literature, and discuss their applicability in the concrete application considered in this work.

Two domain shifts. First note that our task involves two types of domain shift: a domain gap between consumer and shop images within a dataset, and a distribution shift between the source and the target datasets. In this paper we assume that the shift between the consumer and shop domains is more severe than the shift between the two datasets (within the same product categories). However, our proposed approach deals with both simultaneously.

Single dataset case. Let us consider first the case where a single dataset, the one we target, is available. The top 3 rows of Table 1 suggest three scenarios, whose terminology is borrowed from DA for classification: no adaptation (NA), unsupervised domain adaptation (UDA) and supervised domain adaptation (SDA). In the case of SDA, a domain-invariant representation is learned from labeled consumer-shop image pairs. In the case of UDA, however, the consumer product IDs (PIDS) are unknown, so the only solution to tackle this task would be unsupervised adaptation of the feature representation space. If consumer images are unavailable beforehand for training (as in our scenario),

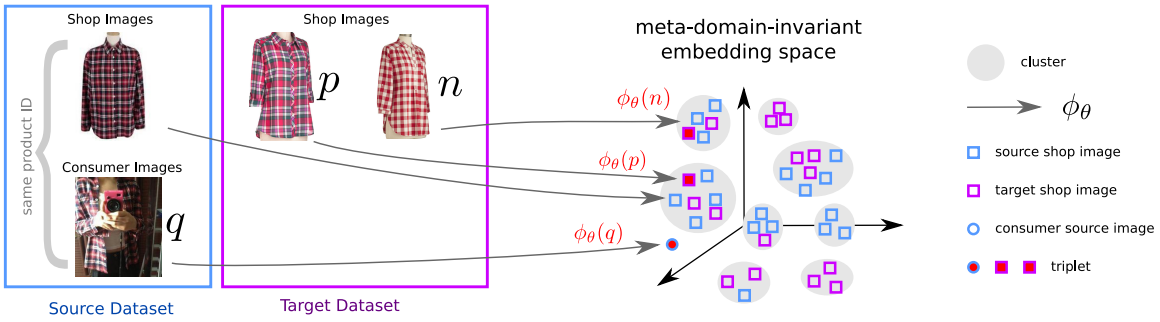

Figure 2. Illustration of our approach: shop images from both the source and target datasets are clustered jointly. Source images within a cluster are used to label the target images. These pseudo-labels can then be used to adapt a source model to the target dataset.

then the system can only learn a representation from the shop images, and has to apply directly this representation to consumer images for C2S retrieval. This approach performs no domain adaptation (NA). A less naive approach, proposed for classification, could be to use single domain generalization (SDG) [49, 37], which augments a domain (the shop domain in our case) in an adversarial manner to make it more resilient to a direct application to unseen domains (such as the consumer domain). Until now, the large body of work on consumer-to-shop (C2S) retrieval has almost exclusively focused on the SDA scenario and a single dataset (see Section 2). This is what we have reviewed in this paragraph. Next we will consider the presence of a second, source, dataset. To the best of our knowledge, this is the first time that a source dataset is used to improve C2S retrieval on a target dataset.

What if a source dataset is available. Let us consider now that we have a source C2S dataset that can be exploited for learning a suitable representation for the target dataset. The easiest possible scenario would be to assume that both the source and the target datasets share product IDs. The corresponding scenarios are presented in the middle part of Table 1. Those describe potential generalizations of DA methods where the adaptation is done across C2S datasets. While we have not seen any proposed method to tackle these scenarios, methods could be devised that assume that the label space is shared between the two datasets (*i.e.* same product IDs). Yet, in our more challenging but more realistic scenario, we can not rely on such a strong simplifying assumption. Instead, we focus on the concrete situation we described above, where the adaptation is performed between two C2S datasets with different sets of products (*e.g.* adaptation of a model trained on Nike products to Adidas ones or between different seasons of Nike products).

Towards meta-domain adaptation. To handle our problem we make the following assumptions. First, we consider that the distribution gap between the source and target datasets is mainly due to the sample bias and more importantly to the fact that the label space (*i.e.* the fashion product IDs) are not shared between datasets. In our experiments, those sets are fully disjoint, but note that if there was a partial overlap of product IDs, this could help with the pseudo-labeling as well as in the training process of our approach. Therefore, instead of trying to solve two domain gaps (between shop and consumer images and between different datasets) we focus on the gap between what we call *meta-domains*, *i.e.* between the meta-domain of all consumer images and the meta-domain of all shop images (see Figure 1), independently of the dataset images come from. Hence, we aim at reducing the gap between these two meta-domains by exploiting a fully supervised source dataset, *i.e.* for which labeled shop-consumer image pairs are available, and some information from the target dataset. Again, we have several cases. The first one simply learns from the source dataset alone, and does not have access to the target dataset: no adaptation other than applying SDA to the source dataset is possible. This is our *unsupervised baseline*. Having zero access to the target domain, we can only hope that the learnt representations generalize well to the target dataset.

Meta-domain adaptation. In our case, we assume access to the target shop images but not to the target consumer images, and we would like to take advantage of these shop images to improve the representation learned on the source dataset so it transfers better to the new target dataset. Additionally, we assume that we have no label information (*i.e.* no product IDs) for the target shop images. Still, note that even if we had access to those, it would remain unsupervised MDA (denoted with an extra * in the table) as we cannot create any ground-truth consumer-shop pairs for the target set. Finally, in the last line of Table 1, we show what would be the supervised meta-domain adaptation case. This is very different from the scenario addressed in this paper.

The next section formalizes unsupervised meta-domain adaptation and describes several models to tackle it.

4. Proposed Approach

Our aim is to train a cross-domain model for consumer-to-shop (C2S) fashion retrieval. We adopt a learning-to-rank approach (section 4.1). It requires relevance labels

between shop and consumer images, unavailable in our scenario. We therefore adapt an existing cross-domain model to a new target dataset, without the need for consumer images (section 4.2). Figure 2 illustrates our approach.

4.1. Learning to Rank

Let $\phi_\theta : I \rightarrow \mathbb{R}^d$ be an embedding function, parameterized by θ, that transforms an image $q \in I$ into a vector representation $\phi(q) \in \mathbb{R}^d$. Learning-to-rank approaches typically produce a function ϕ_θ which embeds relevant images close to each other, so that retrieval can be simply performed by computing and ranking distances in the embedding space. To learn θ, we use the triplet ranking loss:

$$L_{triplet}(\theta) = \sum_{q,p,n} max(0, m + d(\phi_\theta(q), \phi_\theta(p)) - d(\phi_\theta(q), \phi_\theta(n)))$$

where q is a query image, p (resp. n) is a relevant (resp. irrelevant) image for that query, and m is a scalar that controls the margin. Two images are considered relevant to each other if they depict the same fashion item (*e.g*. they share the same product ID). When training with a triplet loss, the triplet sampling strategy is crucial: random sampling yield triplets that incur no loss and therefore do not improve the model. In this work, we follow [14] and use the *batch hard* technique for triplet mining.

Cross-domain consumer-to-shop retrieval. In C2S retrieval, query images from consumers belong to a different domain than shop images in the retailer's database. A typical strategy to bridge this domain gap is to include *cross-domain triplets*, *i.e*. triplets that contain images from both domains. This has been shown to produce domain-invariant embedding functions which work well in practice [19, 24]. In this work we use triplets of the form (q^C, p^S, n^S), (q^C, p^C, n^C) and (q^S, p^S, n^S), where $\cdot^S$ and $\cdot^C$ represent images from the shop and consumer domains respectively. We tried using only triplets of the form (q^C, p^S, n^S) as done in [19] but we observed that using all three types of triplets brought a small but consistent improvement.

4.2. Unsupervised Meta-domain Adaptation

Traditional methods for cross-domain retrieval rely on training with labeled consumer-shop product pairs, but when consumer images are not available beforehand, because *e.g*. a catalog is new, such training is impossible. To tackle this challenging setting, we introduce the concept of *meta-domain* and the problem of *meta-domain adaptation*, which we solve by leveraging a C2S retrieval model learned as described in 4.1 from a *source* dataset.

We denote a *domain* as $\mathcal{D}_G^H$, where $H \in \{C, S\}$ refers to either consumer (C) or shop (S) images, and $G \in \{S, T\}$ to the source (S) or target (T) dataset. For example, $\mathcal{D}_S^S$ is the domain of shop images from a given source dataset. A set of domains $\{\mathcal{D}_\cdot^C\}$ or $\{\mathcal{D}_\cdot^S\}$ constitutes a *meta-domain*. The *meta-domain adaptation* problem can then be defined as the problem of reducing the gap between two meta-domains, *e.g*. consumer and shop meta-domains in our case (see also Figure 1 for an illustration).

Let us consider a consumer $\{\mathcal{D}_S^C, \mathcal{D}_T^C\}$ and a shop $\{\mathcal{D}_S^S, \mathcal{D}_T^S\}$ meta-domain constructed from one source S and one target T dataset. To minimize the meta-domain gap, one could apply the approach described in section 4.1. To do this, one could first form the union of the spaces in each domain and then solve the gap: $\mathcal{D}_S^C \cup \mathcal{D}_T^C \rightarrow \mathcal{D}_S^S \cup \mathcal{D}_T^S$, by constructing (q_U^C, p_U^S, n_U^S) triplets, where $U = S \cup T$ includes both source and target images. In the ideal case, relevance labels would be available between consumer and shop images, and across source and target datasets. This would let us construct triplets of the form (q_S^C, p_T^S, n_T^S) or (q_T^C, p_S^S, n_S^S). Using such triplets, one could learn a domain-invariant, and meta-domain invariant representation, following section 4.1. Note that this scenario would correspond to the Supervised MDA (last row in Table 1).

Unfortunately, we can not apply this approach directly: We cannot build (q_T^C, p_S^S, n_S^S) triplets as we do not have access to $\mathcal{D}_T^C$. Additionally, for (q_S^C, p_T^S, n_T^S) we need shared source and target fashion items. This requirement is generally not verified. Instead, assuming the presence of sufficiently similar products in the two datasets, we propose to propagate product IDs (PIDs) across datasets. We first associate source and target images within the shop meta-domain using a clustering-based approach. Then we propagate the PIDs from source shop instances to target shop instances based on cluster co-occurrences. This pseudo-labeling technique produces relevance labels between target shop images and source consumer image labels, and can be used to generate (q_S^C, p_T^S, n_T^S) triplets for our model adaptation. We describe both of these steps (clustering and pseudo-labeling) in detail in the next paragraphs.

Clustering within the shop meta-domain. To cluster items belonging to different domains in the shop meta-domain, we adopt the FINCH clustering algorithm [45]. FINCH outputs a hierarchical set of partitions that provide a fine-to-coarse view of the produced clustering. FINCH is suitable for our task as it is scalable, does not require specifying the number of clusters, and provides clusters with high purity at finer partitions. Clustering produces three types of clusters: (i) containing only items from $\mathcal{D}_S^S$; (ii) containing only items from $\mathcal{D}_T^S$; and (iii) containing items from both. We retain only clusters of the third type, as pseudo-labeling is only possible within such clusters.

Pseudo-labeling strategies. To propagate the source product IDs (P_i) to target shop images in $\mathcal{D}_T^S$, denoted by T_j, we design several intra-cluster pseudo-labeling strategies. We define several strategies based on combining the following

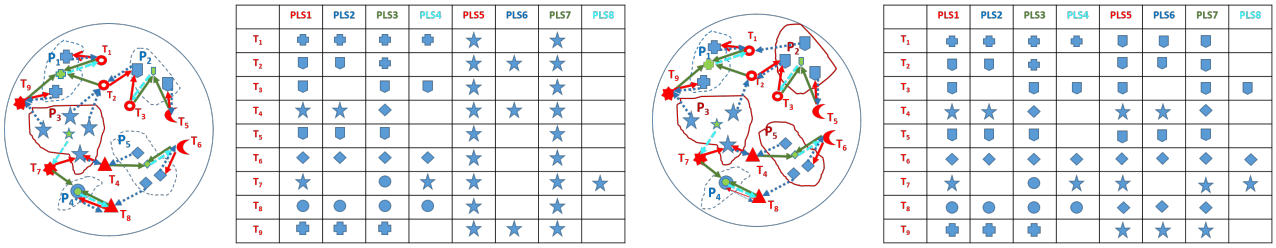

Figure 3. Illustration of different pseudo labeling strategies. Left: single dominant PID. Right: several dominant PIDs.

3 binary choices: (i) mapping target items to source items or the reverse; (ii) maintaining separate image representations for a PID or representing a PID as the average representation of related items; and (iii) propagating labels from all PIDs in a cluster or only from the most dominant PID(s). We compute the dominance of a PID in a cluster as the percentage of items with that PID that are present in the cluster. Computing dominance as the percentage of relevant items in a cluster rather than the frequency selects PIDs that are concentrated in one cluster rather than being dispersed in several different clusters.

These three choices, when combined, lead to eight possible strategies. The **PLS1** strategy labels a target item T_i with the PID of the closest source item. **PLS2** propagates a source item's ID to the closest target item. If several source items are closest to the same target image (see T_2, T_4, T_8 and T_9 as examples in Figure 3), only the source item that is closest to that target item propagates its label to that item. **PLS3/PLS4** are similar to PLS1/PLS2, but average the representations of source items that have the same PID (Figure 3, green smaller shapes). **PLS5/PLS6** are equivalent to PLS1/PLS2 after removing non-dominant PIDs (*e.g.* P_1 and P_4 in Figure 3, right). **PLS7/PLS8** are equivalent to PLS3/PLS4 after removing non-dominant PIDs. All these strategies are illustrated in Figure 3 (see *supplementary* for a higher resolution figure.)

Model adaptation using pseudo-labels. We use the architecture of [13]. We first train it as described in section 4.1 for cross-domain C2S retrieval using DeepFashion [31] as our source dataset. The embedding function obtained after this training step is the one used to extract representations considered by the clustering algorithm. We then refine it with additional (q_S^C, p_T^S, n_T^S) triplets obtained using the pseudo-labeling of the target shop images ($\mathcal{D}_T^S$) provided by one of the strategies described above.

5. Experiments

Section 5.1 describes the used datasets. Section 5.2 gives technical details about the clustering and pseudo-labeling steps. Section 5.3 presents our experimental results.

5.1. Datasets

The **DeepFashion (DF) dataset** [31] is our *source* dataset from which we use both shop and consumer images. For each fashion item, we assume that its category (shirt, dress, etc.) as well as its product ID are known. This allows us to build (consumer, shop) relevance pairs for the different fashion items. DF has four main categories - clothing, dresses, tops and trousers - each composed of subcategories. It contains over 800,000 images annotated with bounding boxes for both consumer and shop images.

The **Street2Shop (S2S) dataset** [21] is our *target* dataset. At train time, we assume access to shop images only, and only use the category labels. This is a weak requirement as one could easily apply a classifier to produce category labels. Product IDs are not used. At test time we evaluate cross-domain retrieval in a standard manner using the S2S test set. S2S has eleven fashion categories: belts, bags, dresses, eyewear, footwear, hats, leggings, outerwear, pants, skirts and tops. The dataset contains 404,683 shop images and 20,357 consumer images. Only consumer images are annotated with bounding boxes. Following the protocol of Kuang *et al.* [23] and Kucer *et al.* [24], we use the cropped shop images available from [24].

Category alignment between datasets. There is no one-to-one correspondence between the categories/subcategories of the two datasets. To overcome this challenge, we establish a mapping between DF subcategories and S2S categories (as shown in the *supplementary material*). We refer to the 6 categories from S2S that we were able to match with DF as *seen categories*. The other five S2S categories, referred as *unseen*, do not match any category in DF and they were not used at train time. Note that not all sub-categories in DF were mapped to the meta-categories.

Other related datasets. There are two additional public cross-modal fashion retrieval datasets one could consider: DeepFashion2 [10] and DARN [17]. However, DeepFashion2 is simply an extension of DF. As for DARN, only a small subset of its images contain bounding box annotations while most images contain multiple items. Working with full-body images containing multiple items is a different problem and requires a different pipeline. Therefore, we conduct all experiments using DF and S2S.

Evaluation. Following the standard S2S [21] protocol, we report top-k accuracy and mean Average Precision (mAP).

5.2. Implementation Details

Baseline. Using the DF dataset, we pre-train the model described in section 4.1 using (q_S^S, p_S^S, n_S^S), (q_S^C, p_S^C, n_S^C) and (q_S^C, p_S^S, n_S^S) triplets, which mix consumer (C) and shop (S) samples. This model is our main baseline, (first row in the third part of Table 1), and we refer to it as `DF-BL`. All our proposed models rely on this baseline model, either using it as a feature extractor, or fine-tuning it using pseudo-labels. The joint clustering of DF and S2S shop images is also done using representations extracted from `DF-BL`.

We use the `Resnet101-TL-GeM` architectures from [13] for `DF-BL` and, consequently, all our subsequent models. We train `DF-BL` using both the train and validation sets of DF in order to obtain the strongest possible feature extractor. `DF-BL` achieves 67.1% top-20 accuracy on the DF test set (seen and unseen categories). We also include results with ResNet50 backbone (63.59% top-20 accuracy on the DF test set) for comparison.

Meta-domain adaptation (`UMDA`). We adapt the baseline model using our proposed unsupervised meta-domain (or `UMDA`) approach by fine-tuning it using, in addition to the three triplet types described in the baseline model, (q_S^C, p_T^S, n_T^S) triplets, where the consumer image q_S^C is from DF and the shop images p_T^S and n_T^S are from S2S. Relevance labels come from one of the pseudo-labeling strategies (see section 4.2). As the fashion category of each item is assumed to be known, we only form such triplets with items from the same category. For the same reason, we do not cluster all shop images from DF and S2S together, but perform one independent clustering for each meta-category, ensuring better clustering and label propagation. We investigate two types of fine-tuning: shallow fine-tuning (`UMDA-MLP`), where we only fine-tune the last fully-connected layer of the network, and end-to-end fine-tuning (`UMDA-E2E`), where we fine-tune all network layers.

Training and optimization. Our baseline and `UMDA` models are implemented in PyTorch v1.4. We optimize the model parameters using Adam [22]. We initialize the learning rate to 10^{-4} and decrease it by a factor of 10 every 15,000 iterations. The maximum number of iterations is set to 45,000. We use a batch size of 128. We use the above training procedure for both `UMDA-MLP` and `UMDA-E2E` models. Furthermore, for `E2E` model training, we perform standard image augmentation with random horizontal flips and rotations (45 degrees). All images are resized to 256×256, randomly cropped to 224×224 pixels, and then mean-subtracted for network training.

Clustering. The FINCH algorithm produces multiple partitions. The first one corresponds to linking samples through the first neighbor relations, while the second one links clusters created in the first step. We use clusters from the first partition to mine positive and negative pairs as this partition provides the best compromise between diversity and quality. Further, we show some qualitative results in Figure 4. Further qualitative examples as well as statistics of the joint DF and S2S shop image clusters are shown in the *Supplementary material*. From these results, one can indeed see diverse but semantically and stylistically coherent clusters.

5.3. Quantitative Results

Pseudo-labeling strategies. We first compare several pseudo-labeling strategies using the `UMDA-MLP` model in Table 3. We do not report PLS4 and PLS8 as they lead to too few triplets to train properly. We report PLS7, but not PLS3 and PLS5, as PLS7 already includes the modifications of PLS3 and PLS5 with respect to PLS1, our most basic strategy. As Table 3 shows, PLS6 and PLS7 perform best, with PLS7 being slightly better. Consequently, we use the PLS7 strategy for all further experiments. This means that S2S shop images are assigned the dominant DF fashion PID in the cluster. If multiple PIDs have the same dominance score, each S2S shop image is assigned the closest averaged PID representation.

Comparisons to the `DF-BL` baseline. Table 4 compares our `UMDA-MLP` and `UMDA-E2E` models to the `DF-BL` baseline for two backbones - ResNet-50 and ResNet-101. We compare using mAP and Top-$\{$1,5,10,20,50$\}$ metrics. We can see that for both backbones, our `UMDA-MLP` approach results in consistent performance improvements for all metrics. Our `UMDA-E2E` provides further performance improvements over `UMDA-MLP`, also consistently across all metrics and for both backbones. In particular, `UMDA-E2E` achieves an absolute improvement of 2.49% compared with `DF-BL`, with ResNet-101. This illustrates the effectiveness of our cluster-based unsupervised meta-domain adaptation approach. We can see that both `UMDA` approaches significantly improve upon `DF-BL`. Furthermore, we observe similar behavior on unseen categories (as shown in the *Supplementary material*).

Comparisons to state-of-the-art methods. As none of the methods in the literature tackles unsupervised meta-domain adaptation or considers a scenario similar to ours, we can only put our model's performance in perspective by looking at methods trained in a fully supervised manner on the S2S training set. A direct comparison, however, is not possible because many of these published methods have different underlying architectures and training schemes.

In Table 2 we compare our two models with existing supervised state-of-the-art methods. These include models trained independently for each category [21], and methods that use a common model trained for all categories at once [51, 24, 23]. We note that our supervised base-

Figure 4. Example images from two clusters (from clustering DF and S2S shop images) and an associated DF consumer-DF/S2S shop image pairing (red arrow) example, using one of the dominant PID (images with dotted blue frame, strategy PLS7). Best viewed in color.

Table 2. Comparison with the state of the art. Top-20 accuracy for cross-modal retrieval on S2S. Note that both scenarios are not directly comparable (see text for details).

Scenario	Approach	Backbone	Dresses	Leggings	Outerwear	Pants	Skirts	Tops	Average
Supervised	Kiapour *et al.* [21]	AlexNet	37.1	22.1	21.0	29.2	54.6	54.6	33.7
	GRNet [23]	GoogleNet	64.2	-	38.6	48.5	72.5	58.3	-
	Wang *et al.* [51]	BN-Inception	56.9	15.9	20.3	22.3	50.8	48.0	35.7
	Kucer *et al.* [24]	ResNet-50	**71.6**	**47.3**	44.8	50.0	79.4	59.3	58.7
	Our baseline	ResNet-50	63.31	43.27	43.61	46.67	78.85	54.08	54.96
	Our baseline	ResNet-101	67.12	43.27	46.89	48.33	**82.97**	56.05	57.44
Unsupervised	DF-BL	ResNet-50	59.57	32.65	37.38	51.67	75.55	54.90	51.95
	UMDA-MLP	ResNet-50	60.47	31.43	38.03	51.67	76.10	54.41	52.02
	UMDA-E2E	ResNet-50	62.34	33.29	39.45	56.54	77.92	56.11	54.44
Unsupervised	DF-BL	ResNet-101	60.58	38.21	40.33	50.00	75.55	58.17	53.81
	UMDA-MLP	ResNet-101	63.00	39.43	43.61	56.67	76.65	58.99	56.39
	UMDA-E2E	ResNet-101	67.00	41.89	**45.21**	**63.05**	79.01	**62.40**	**59.76**

Table 3. Impact of the choice of the pseudo-labeling strategy on cross-modal retrieval on seen 6 categories of S2S.

	mAP	Top-1	Top-5	Top-10	Top-20
PLS1	23.79	35.22	47.64	51.82	55.63
PLS2	23.86	35.31	47.85	52.19	55.51
PLS6	24.30	**36.57**	46.95	51.87	55.72
PLS7	**24.51**	36.12	**47.74**	**52.70**	**56.39**

Table 4. Comparison of Baseline models ResNet-50/101

		mAP	Top-1	Top-5	Top-10	Top-20	Top-50
		Seen (6 Categories)					
Res50	DF-BL	22.60	32.75	43.56	47.47	51.95	59.38
	UDMA-MLP	23.65	34.71	44.77	48.11	52.02	60.06
	UDMA-E2E	24.26	35.89	45.96	49.88	54.44	62.35
Res101	DF-BL	22.97	32.97	45.08	49.21	53.81	59.39
	UDMA-MLP	24.51	36.12	47.74	52.70	56.39	62.40
	UDMA-E2E	**25.46**	**37.80**	**49.68**	**55.13**	**59.76**	**65.18**

line trained on ResNet-50 gives lower performance than the method of Kucer *et al.* [24]. We believe this is due to the more sophisticated data augmentations and the more computationally-expensive training regimes (e.g. using much higher batch sizes) used by Kucer *et al.* [24]. We expect that all of our baselines and proposed models can be improved using such regimes.

Nevertheless, our unsupervised UMDA-E2E models perform quite on par with these supervised methods. Further, our best unsupervised model, UMDA-E2E with ResNet-101, improves over the previous state-of-the-art supervised approach [24] for Outerwear, Skirts, Top, and on average. With the comparable ResNet-50 architecture we achieve better or on-par performance compared to these fully supervised methods albeit having been trained in a much more challenging unsupervised meta-domain adaptation setting.

6. Conclusion

In this work we tackle a challenging yet realistic cross-domain fashion item retrieval scenario where, in contrast to existing approaches, we assume that we have no access to consumer images because, for instance, the service is newly deployed and no user query is available yet. To address this problem, we propose an unsupervised meta-domain adaptation method that relies on a clustering-based pseudo-labeling strategy to leverage cross-domain representations learned from an existing labeled consumer-to-shop dataset. We experimentally show that the proposed approach achieves good cross-modal retrieval performance on the Street-to-Shop benchmark, without having access to any consumer images of this dataset at train time.

References

[1] Sławomir Bak, Peter Carr, and Jean-François Lalonde. Domain Adaptation through Synthesis for Unsupervised Person Re-identification. In *Proc. ECCV*, 2018.

[2] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial Adversarial Domain Adaptation. In *Proc. ECCV*, 2018.

[3] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to Transfer Examples for Partial Domain Adaptation. In *Proc. CVPR*, 2019.

[4] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-Specific Batch Normalization for Unsupervised Domain Adaptation. In *Proc. CVPR*, 2019.

[5] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive Feature Alignment for Unsupervised Domain Adaptation. In *Proc. CVPR*, 2019.

[6] Gabriela Csurka. A Comprehensive Survey on Domain Adaptation for Visual Applications. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 1–35. Springer, 2017.

[7] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In *Proc. ECCV*, 2018.

[8] Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *Proc. CVPR Workshop*, 2018.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-Adversarial Training of Neural Networks. *JMLR*, 2016.

[10] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5337–5345, 2019.

[11] Raghuraman Gopalan, Ruonan Li, and Vishal M. Patel. Foundations and Trends in Computer Graphics and Vision. Now Publishers Inc., 2015.

[12] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. ECCV*, 2016.

[13] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 2017.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrel. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proc. ICML*, 2018.

[16] Fuxiang Huang, Lei Zhang, Yang Yang, and Xichuan Zhou. Probability weighted compact feature for domain adaptive retrieval. In *Proc. CVPR*, 2020.

[17] Junshi Huang, Rogerio Feris, Qiang Chen, and Shuicheng Yan. Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network. In *Proc. ICCV*, Santiago, Chile, 2015.

[18] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *Proc. MM*, 2017.

[19] Shuhui Jiang, Yue Wu, and Yun Fu. Deep bi-directional cross-triplet embedding for cross-domain clothing retrieval. In *Proc. MM*, 2016.

[20] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive Adaptation Network for Unsupervised Domain Adaptation. In *Proc. CVPR*, 2019.

[21] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to Buy It: Matching Street Clothing Photos in Online Shops. In *Proc. ICCV*, 2015.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, (abs/1412.6980), 2014.

[23] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proc. CVPR*, pages 3066–3075, 2019.

[24] Michal Kucer and Naila Murray. A detect-then-retrieve model for multi-domain fashion item retrieval. In *FFSS-USAD CVPR Workshop*, 2019.

[25] Jian Liang, Yunbo Wang, Dapeng Hu, Ran He, and Jiashi Feng. A Balanced and Uncertainty-aware Approach for Partial Domain Adaptation. In *Proc. ECCV*, 2020.

[26] Anan Liu, Shu Xiang, Wenhui Li, Weizhi Nie, and Yuting Su. Cross-Domain 3D Model Retrieval via Visual Domain Adaptation. In *Proc. IJCAI*, 2018.

[27] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proc. CVPR*, 2019.

[28] Ming-Yu Liu and Oncel Tuzel. Coupled Generative Adversarial Networks. In *Proc. NIPS*, 2016.

[29] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proc. CVPR*, 2012.

[30] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2017.

[31] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR*, 2016.

[32] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion Landmark Detection in the Wild. In *Proc. ECCV*, 2016.

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. In *Proc. ICML*, 2015.

[34] Andy Jinhua Ma, Jiawei Li, Pong C. Yuen, and Ping Li. Cross-Domain Person Reidentification Using Domain Adaptation Ranking SVMs. *TIP*, 24(5):1599–1613, 2015.

[35] Pau Panareda Busto and Juergen Gall. Open Set Domain Adaptation. In *Proc. ICCV*, 2017.

[36] Yuxin Peng and Jingze Chi. Unsupervised Cross-media Retrieval Using Domain Adaptation with Scene Graph. *Trans. CSVT*, early access, 2019.

[37] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to Learn Single Domain Generalization. In *Proc. CVPR*, 2020.

[38] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Proc. ECCV*, 2016.

[39] Jerome Revaud, Jon Almazan, Rafael S. Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proc. ICCV*, 2019.

[40] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Residual Parameter Transfer for Deep Domain Adaptation. In *Proc. CVPR*, 2018.

[41] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial Dropout Regularization. In *Proc. ICLR*, 2018.

[42] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open Set Domain Adaptation by Backpropagation. In *Proc. ECCV*, 2018.

[43] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood reranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[44] M Saquib Sarfraz and M Haris Khan. A probabilistic framework for patch based vehicle type recognition. In *Visapp*, 2011.

[45] M. Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proc. CVPR*, 2019.

[46] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning Transferrable Representations for Unsupervised Domain Adaptation. In *Proc. NIPS*, 2016.

[47] Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Proc. ECCV Workshop*, 2016.

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *Proc. CVPR*, 2017.

[49] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to Unseen Domains via Adversarial Data Augmentation. In *Proc. NIPS*, 2018.

[50] Shanshan Wang and Lei Zhang. Self-adaptive re-weighted adversarial domain adaptation. In *Proc. IJCAI*, 2020.

[51] Xi Wang, Zhenfeng Sun, Wenqiang Zhang, Yu Zhou, and Yu-Gang Jiang. Matching User Photos to Online Products with Robust Deep Features. In *Proc. ICMR*, 2016.

[52] Zhonghao Wang, Yujun Gu, Ya Zhang, Jun Zhou, and Xiao Gu. Clothing retrieval with visual attention model. In *IEEE Visual Communications and Image Processing (VCIP)*, 2017.

[53] Yichao Xiong, Ning Liu, Zhe Xu, and Ya Zhang. A parameter partial-sharing cnn architecture for cross-domain clothing retrieval. In *IEEE Visual Communications and Image Processing (VCIP)*, 2016.

[54] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and Adversarial Network for Unsupervised Domain Adaptation. In *Proc. CVPR*, 2018.

[55] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual search at alibaba. In *Proc. KDD*, 2018.

[56] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification. In *Proc. CVPR*, 2019.