Where to Buy It: Matching Street Clothing Photos in Online Shops

M. Hadi Kiapour[1], Xufeng Han[1], Svetlana Lazebnik[2], Alexander C. Berg[1], Tamara L. Berg[1]

[1]University of North Carolina at Chapel Hill
{hadi,xufeng,tlberg,aberg}@cs.unc.edu
[2]University of Illinois at Urbana-Champaign
slazebni@illinois.edu

Abstract

In this paper, we define a new task, Exact Street to Shop, where our goal is to match a real-world example of a garment item to the same item in an online shop. This is an extremely challenging task due to visual differences between street photos (pictures of people wearing clothing in everyday uncontrolled settings) and online shop photos (pictures of clothing items on people, mannequins, or in isolation, captured by professionals in more controlled settings). We collect a new dataset for this application containing 404,683 shop photos collected from 25 different online retailers and 20,357 street photos, providing a total of 39,479 clothing item matches between street and shop photos. We develop three different methods for Exact Street to Shop retrieval, including two deep learning baseline methods, and a method to learn a similarity measure between the street and shop domains. Experiments demonstrate that our learned similarity significantly outperforms our baselines that use existing deep learning based representations.

1. Introduction

Online shopping is an exponentially growing market. Retail sales world-wide, including both in-store and internet purchases, totaled approximately $22.5 trillion in 2014, with $1.316 trillion of sales occurring online. By 2018, ecommerce retail spending is projected to increase to nearly $2.5 trillion[1]. Much of this purchasing is related to shopping for clothing items. However, finding exactly what you want from online shops is still not a solved problem.

In this paper, we look at one task related to online shopping, the street-to-shop problem. Given a real-world photo of a clothing item, e.g. taken on the street, the goal of this task is to find that clothing item in an online shop. This is extremely challenging due to differences between depictions of clothing in real-world settings versus the clean simplicity of online shopping images. For example, clothing will be worn on a person in street photos, whereas in online shops, clothing items may also be portrayed in isolation or on mannequins. Shop images are professionally photographed, with cleaner backgrounds, better lighting, and more distinctive poses than may be found in real-world, consumer-captured photos of garments. To deal with these challenges, we introduce a deep learning based methodology to learn a similarity measure between street and shop photos.

Figure 1: Our task is to find the exact clothing item, here a dress, shown in the query. Only the first dress, in the green rectangle, would be considered correct. This is different from previous work, e.g. [24], that considers whether retrieved items have similar high-level features. Under that, more relaxed, evaluation all of the dresses shown are correct. (For this query, our similarity learning ranked the correct match first.)

The street-to-shop problem has been recently explored [24]. Previously, the goal was to find *similar* clothing items in online shops, where performance is measured according to how well retrieved images match a fixed set of attributes, e.g. color, length, material, that have been hand-labeled on the query clothing items. However, finding a similar garment item may not always correspond to what a shopper desires. Often when a shopper wants to find an item online, they want to find *exactly* that item to purchase.

Therefore, we define a new task, *Exact Street to Shop*, where our goal is for a query street garment item, to find exactly the same garment in online shopping images (Fig. 1).

[1]http://www.emarketer.com/Article/Retail-Sales-Worldwide-Will-Top-22-Trillion-This-Year/1011765

To study Exact Street to Shop at large scale, we collected and labeled a dataset of 20,357 images of clothing worn by people in the real world, and 404,683 images of clothing from shopping websites. The dataset contains 39,479 pairs of exactly matching items worn in street photos and shown in shop images. While small relative to all shopping images on the web, we have gone far past the "Dress Barn" (a clothing chain in the US) and are working at the scale of a "Dress Aircraft Hanger"!

Our paper attacks the Exact Street to Shop problem using multiple methods. We first look at how well standard deep feature representations on whole images or on object proposals can perform on this retrieval task. Then, we explore methods to learn similarity metrics between street and shop item photos. These similarities are learned between existing deep feature representations extracted from images. To examine the difficulty of the Exact Street to Shop task and to evaluate our retrieval results, we also provide several human experiments, evaluating when and where exact item retrieval is feasible.

In summary, our contributions are:

- Introduction of the Exact Street to Shop task and collection of a novel dataset, the Exact Street2Shop Dataset, for evaluating performance on this task.
- Development and evaluation of deep learning feature based retrieval and similarity learning methods for the Exact Street to Shop retrieval task.
- Human evaluations of the Exact Street to Shop task and of our results.

The rest of our paper is organized as follows: First, we review some related works (Sec. 2). Next, we describe our new dataset (Sec. 3) and approaches (Sec. 4) for the Exact Street to Shop task. Finally, we provide experimental results (Sec. 5) and conclusions (Sec. 6).

2. Related Work

Clothing Recognition: There has been growing interest in clothing recognition from the computer vision and multimedia communities. Some recent papers have demonstrated effective methods for clothing parsing, where the goal is to assign a semantic clothing label to each pixel in an image of a person [35, 38, 37, 7, 22]. Other works have explored ways to identify aspects of a person's socio-identity, including predicting their social tribe [19, 17], fashionability [36, 30], or occupation [31] from the clothing they are wearing. Several methods have used attribute-based frameworks to describe, classify, or retrieve clothing [3, 4, 5].

Image Retrieval: Image retrieval is a fundamental problem for computer vision with wide applicability to commercial systems. Many recent retrieval methods at a high-level consist of three main steps: pooling local image descriptors (such as Fisher Vectors [25, 27, 26] or VLAD [15]), dimensionality reduction, and indexing. Lim et al. [21] used keypoint detectors to identify furniture items by aligning 3D models to 2D image regions. Generally, these methods work quite well for instance retrieval of rigid objects, but may be less applicable for retrieving the soft, deformable clothing items that are our focus.

Clothing Retrieval: Despite recent advances in generic image retrieval, there have been relatively few studies focused specifically on clothing retrieval. Some related works have performed garment retrieval using parsing [38], or using global or fine-grained attribute prediction [5]. There have also been some efforts on cross-scenario retrieval [24, 23, 9, 16]. Most related to our work is the street-to-shop [24] approach, which tackles the domain discrepancy between street photos and shop photos using sparse representations. However, their approach depends on upper/lower body detectors to align local body parts in street and shop images, which may not be feasible in all types of shop images. They also evaluate retrieval performance in terms of a fixed set of hand-labeled attributes. For example, evaluating whether both the query and shop images depict a "blue, long-sleeved, shirt". While this type of evaluation may suit some shoppers' needs, we posit that often a shopper's goal is to find *exactly* the same street item in an online shop.

Deep Similarity: As deep convolutional neural networks are becoming ubiquitous for feature representations, there has been growing interest in similarity learning with deep models. Some examples include methods for fine-grained object retrieval [34, 20], face verification [29, 32], or image patch-matching [40, 13, 41]. These techniques learn representations coupled with either predefined distance functions, or with more generic learned multi-layer network similarity measures. For our similarity learning method, we learn a multi-layer network similarity measure on top of existing pre-trained deep features.

Domain Adaptation: The concept of adapting models between different dataset domains has been well explored. Many works in this area tackle the domain adaptation problem by learning a transformation that aligns the source and target domain representations into a common feature space [1, 8, 12, 11]. Other approaches have examined domain adaptation methods for situations where only a limited amount of labeled data is available in the target domain. These methods train classifiers on the source domain and regularize them against the target domain [2, 28]. Recently, supervised deep CNNs have proved to be extremely successful for the domain adaptation task [6, 14, 39]. Our data can be seen as consisting of two visual domains, shop images and street images. In contrast to most domain adaptation techniques that try to adapt for a classification task, our method retrieves items across domains.

Figure 2: Example *street outfit photos*, including large variations in pose, camera angle, composition and quality.

Figure 3: Example *shop photos*, displaying a wide range of apparel photography techniques.

3. Dataset

We collect a novel dataset, the *Exact Street2Shop Dataset*, to enable retrieval applications between real world photos and online shopping images of clothing items. This dataset is available online at http://www.tamaraberg.com/street2shop. It contains two types of images: *1) street photos*, which are real-world photographs of people wearing clothing items, captured in everyday uncontrolled settings, and *2) shop photos*, which are photographs of clothing items from online clothing stores, worn by people, mannequins, or in isolation, and captured by professionals in more controlled settings. Particular clothing items that occur in both the street and shop photo collections form our *exact street-to-shop pairs* for evaluating retrieval algorithms. In the following sections we describe our dataset and annotation process in detail.

3.1. Image Collection

In this section, we describe our data collection of street photos (Sec. 3.1), shop photos (Sec. 3.1.2), and correspondences between street and shop items (Sec. 3.1.3).

3.1.1 Street Photos

To create a useful dataset for evaluating clothing retrieval algorithms, we would like to collect street photographs of clothing items for which we know the correspondences to the same clothing items in online shops. There are a number of social communities focused on fashion, such as Chictopia.com and various fashion blogs, where people post photographs of themselves wearing clothing along with links to purchase the items they are wearing. However, these links are often outdated or point to items that are similar but not identical to the items being worn.

Instead, to gather corresponding street-shop item pairs for a wide range of different people and environments, we make use of style galleries from ModCloth[2]. ModCloth is a large online retail store specializing in vintage style fashion that sells clothes from a wide variety of brands. These style galleries contain user-contributed outfit posts, in which people upload photos of themselves wearing ModCloth clothing items and provide shopping links to the exact items they are wearing.

We collect 20,357 style gallery outfit posts, spanning user-contributed photos (example outfit photos are shown in Fig. 2). Each outfit post consists of a *street photo* that depicts at least one of the clothing items offered on the ModCloth website. These photographs aim to showcase how one would style an outfit and help others decide whether they want to purchase an item. There are large variations in the quality of the contributed photographs, lighting, indoor vs outdoor environments, body shapes and sizes of the people wearing the clothing, depicted pose, camera viewing angle, and a huge amount of occlusion due to layering of items in outfits. In addition, a photo may depict a head-to-toe shot or several partial-body shots. These characteristics reflect the extreme challenges and variations that we expect to find for clothing retrieval in real-world applications.

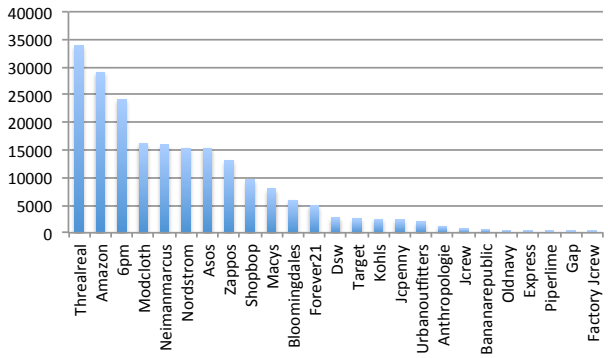
3.1.2 Shop Photos

We have collected 404,683 shop photos from 25 different online clothing retailers. These photos depict 204,795 distinct clothing items (each clothing item may be associated with multiple photographs showing different views of the item). Moreover, when available, the title and a detailed description of the item is extracted from the product's webpage. We collect 11 different broad categories of clothing items (See Table 2), ranging from small items such as belts and eyewear, to medium size items such as hats or footwear, to larger items such as dresses, skirts, or pants.

Shop photos differ drastically from street photos in that they are professionally produced and tend to be high-resolution with clean backgrounds, captured under nice lighting with highly-controlled conditions. Different brands

[2]http://www.ModCloth.com

Figure 4: Distribution of collected items across shopping sites.



have different styles of fashion photography, ranging from more basic depictions to professional models. In addition, while some shop photos display a clothing product on a full or partial mannequin, or on a live model, others depict clothing items folded or lying flat on a surface. Shop images also often include close-up shots that display clothing details such as fabric texture or pattern. Altogether, these qualities make our shop dataset highly diverse. Example shop photos are shown in Fig. 3, and the distribution of collected items across shopping websites in our dataset is displayed in Fig. 4.

3.1.3 Street-to-Shop Pairs

Each street photo in our dataset is associated with two types of links to shop clothing items: the first set contains links to products that *exactly* match one of the pictured items in a street photo, while links in the second set indicate items that are only *similar* to a street item. These links are user-provided, but we have manually verified that the links are highly accurate. We make use of only the exact matching items to create our street-to-shop pairs, but we also release the similar matching pairs in our public dataset for evaluation of other types of image retrieval algorithms. In total, there are 39,479 exact matching street-to-shop item pairs.

3.2. Image Annotation

For the retrieval task, we assume that we know two things about a query street image: 1) what category of item we are looking for, and 2) the location of the item in the image. In a real-world retrieval application, this information could easily be provided by a motivated user through input of a bounding box around the item of interest and selection of a high-level category, e.g. skirt. Therefore, we pre-annotate our dataset in two ways. First, we automatically associate a high-level garment category with each item in the dataset (Sec. 3.2.1). Then, we collect bounding boxes for each street query item (Sec. 3.2.2). The latter task is performed using Amazon's Mechanical Turk service.

Category	Keywords
bags	backpack, backpacks, bag, bags, clutch, clutches, evening-handbags, hobo-bag, hobo-bags, satchel, satchels, shoulder-bags, tote-bags, wallet, wallets
dresses	bridal-dresses, bridal-mother, bride, bridesmaid, casual-dresses, cocktail-dresses, day-dresses, dress, dress-pants, evening-dresses, fit-flare-dresses, gown, gowns, longer-length-dresses, maternity-dresses, maxi-dresses, party-dresses, petite-dresses, plus-size-dresses, special-occasion-dresses, teen-girls-dresses, work-dresses
eyewear	glasses, sunglasses, womens-eyewear
footwear	boot, boots, evening-shoes, flats, heel, mules-and-clogs, platforms, pump, pumps, sandal, sandals, shoe, shoes-athletic, shoes-boots, shoes-flats, shoes-heels, shoes-sandals, shoes-wedges, slipper, slippers, wedges, womens-sneakers

Table 1: Example mappings between keywords and high-level item categories.

3.2.1 Category Labeling

Our category labeling on the shop side relies on the metadata associated with collected items. For every item, its product category on the website, web url, title, and description are collected if available. We then create a mapping between product keywords and our final list of 11 garment categories: bags, belts, dresses, eyewear, footwear, hats, leggings, outerwear, pants, skirts, and tops. Finally, we label every item with the category associated with the keywords found in its metadata. A sample of the mappings from keywords to garment categories is shown in Table 1.

3.2.2 Instance Annotation

For every street-to-shop pair, we collect bounding boxes for the item of interest in the street photograph. Note, street photos depict entire outfits, making it necessary to present both the street photo and the corresponding shop photos to the Turker during this annotation process. In particular, we show the workers a street photo and the corresponding shop item photos and ask the Turker to annotate instances of the shop item in the street photograph. Annotators draw a tight bounding box around each instance of the shop item in the provided street photo. To aid this process Turkers are provided with example annotations for each item type, including items with multiple objects, e.g. pairs of shoes.

4. Approaches

We implement several different retrieval methods for the street-to-shop matching problem. Inputs to our methods are a street query image, the category of item of interest, and a bounding box around the item in the query image. On the shop side, since there are a large number of images, we do not assume any hand-labeled localization of items, instead letting the algorithm rely on features computed on the entire image or on object proposal regions.

We first present two baseline retrieval methods for the street-to-shop task using deep learning features as descrip-

tors for matching to an entire shop image (Sec. 4.1) or to object proposal regions within the shop images (Sec. 4.2). Next, we describe our approach to learn a similarity metric between street and shop items using deep networks (Sec. 4.3).

4.1. Whole Image Retrieval

In this approach, we apply the widely used CNN model of Krizhevsky et al. [18], pre-trained for image classification of 1000 object categories on ImageNet. As our feature representation, we use the activations of the fully-connected layer FC6 (4096 dimensions). For query street photos, since we have item bounding boxes available, we compute features only on the cropped item region. For shop images, we compute CNN features on the entire image. We then compare the cosine similarity between the query features and all shop image features and rank shop retrievals based on this similarity.

4.2. Object Proposal Retrieval

In this approach, we use the selective search method [33] to extract a set of object proposals from shop photos. Ideally, the proposed windows will encapsulate visual signals from the clothing item, limiting the effects of background regions and leading to more accurate retrievals. In addition, this step should serve to reduce some of the variability observed across different online shops and item depictions.

Specifically, we use the selective search algorithm and filter out any proposals with a width smaller than $\frac{1}{5}$ of the image width since these usually correspond to false positive proposals. From this set, the 100 most confident object proposals are kept. This remaining set of object proposals has an average recall of 97.76%, evaluated on an annotated subset of 13,004 shop item photos. Similar to the whole image retrieval method, we compute FC6 features on the street item bounding box and on the 100 most confident object proposals for each shop image. We then rank shop item retrievals using cosine similarity.

4.3. Similarity Learning

In this approach, our goal is to learn a similarity measure between query and shop items. Our hypothesis is that the cosine similarity on existing CNN features may be too general to capture the underlying differences between the street and shop domains. Therefore, we explore methods to learn the similarity measure between CNN features in the street and shop domains.

Inspired by recent work on deep similarity learning for matching image patches between images of the same scene [13, 40, 41], we model the similarity between a query feature descriptor and a shop feature descriptor with a three-layer fully-connected network and learn the similarity parameters for this architecture. Here, labeled data for training consists of positive samples, selected from exact street-to-shop pairs, and negative samples, selected from non-matching street-to-shop items.

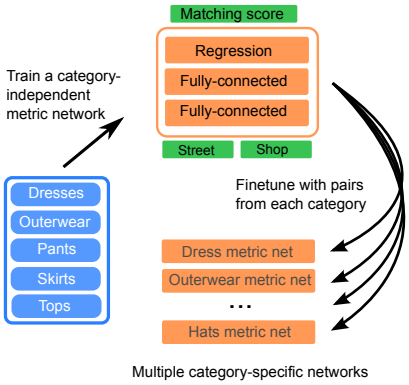


Figure 5: Illustration of the training, followed by fine-tuning procedure for training category-specific similarity for each category. To deal with limited data, we first train a generic similarity using five large categories and then fine-tune it for each category individually. See Sec. 4.3 for more description.

Specifically, the first two fully-connected layers of our similarity network have 512 outputs and use Rectified Linear Unit (ReLU) as their non-linear activation function. The third layer of our network has two output nodes and uses the soft-max function as its activation function. The two outputs from this final layer can be interpreted as estimates of the probability that a street and shop item "match", or "do not match", which is consistent with the use of cross-entropy loss during training. Once we have trained our network, during the test phase, we use the "match" output prediction as our similarity score. Previous work has shown that this type of metric network has the capacity for approximating the underlying non-linear similarity between features. For example, Han et al. [13] showed that the learned similarity for SIFT features, modeled by such a network, is more effective than L2-distance or cosine-similarity for matching patches across images of a scene.

We formulate the similarity learning task as a binary classification problem, in which positive/negative examples are pairs of CNN features from a query bounding box and a shop image selective-search based item proposal, for the same item/different items. We minimize the cross-entropy error

$$E = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \qquad (1)$$

over a training set of n bounding box pairs using mini-batch stochastic gradient descent. Here, $y_i = 1$ for positive examples; $y_i = 0$ for negative examples; and $\hat{y}_i$ and $1 - \hat{y}_i$ are the two outputs of the metric network. One complication is that we do not have hand-labeled bounding boxes for shop images. We could use all object proposals for a shop image in a matching street-to-shop pair as posi-

Figure 6: Example retrievals. Top and bottom three rows show example successful and failure cases respectively.

tive training data, but because many boxes returned by the selective-search procedure will have low intersection-over-union (IoU) with the shop item of interest, it would introduce too many noisy training examples. Another source of noisy examples for similarity training is that, due to large pose differences in images for an item, some images on the shop side will bear little similarity in appearance to a particular query item view. Labeling such visually distinct pairs as positives would likely confuse the classifier during training.

We handle these challenges by training our metric network on a short list of top retrieved shop bounding boxes using the object proposal retrieval approach described in Sec. 4.2. At test time, we similarly use the object proposal approach to provide a short list of candidate retrievals and then re-rank this list using our learned similarity. This has an added benefit of improving the efficiency of our retrieval approach since the original cosine similarity measure is faster to compute than the learned similarity.

More specifically, to construct training and validation sets for similarity learning, for each training query item, q, we retrieve the top 1000 selective search boxes from shop images using cosine similarity. For each bounding box b from a shop image in this set, (q, b) is a positive sample if the shop image is a street-to-shop pair with q. Otherwise, (q, b) is used as a negative sample[3].

Intuitively, we might want to train a different similarity measure for each garment category, for example, objects such as hats might undergo different deformations and transformations than objects like dresses. However, we are limited in the number of positive training examples for each category and by the large negative-to-positive ratio. Therefore, we employ negative sampling to balance the positive and negative examples in each mini-batch. We train a general street-to-shop similarity measure, followed by fine-tuning for each garment category to achieve *category-specific* similarity (See Fig. 5).

In the first stage of training, we select five large categories from our garment categories: Dresses, Outerwear, Pants, Skirts, and Tops and combine their training examples. Using these examples, we train an initial *category-*

[3]Note, here we use only shop bounding boxes for training belonging to the top-K ($K = 75$) items in the retrieval set

independent metric network. We set the learning rate to 0.001, momentum to 0.9, and train for 24,000 iterations, then lower the learning rate to 0.0001 and train for another 18,000 iterations. In the second stage of learning, we fine-tune the learned metric network on each category independently (with learning rate 0.0001), to produce category-dependent similarity measures. In both stages of learning, the corresponding validation sets are used for monitoring purposes to determine when to stop training.

5. Experimental Results

The proposed retrieval approaches are evaluated with a series of retrieval experiments. For these experiments, we split the exact matching pairs into two disjoint sets such that there is no overlap of items in street and shop photos between train and test. In particular, for each category, the street-to-shop pairs are distributed into train and test splits with a ratio of approximately 4:1. For our retrieval experiments, a query consists of two parts: 1) a street photo with an annotated bounding box indicating the target item, and 2) the category label of the target item. We view these as simple annotations that a motivated user could easily provide, but this could be generalized to use automatic detection methods. Since the category is assumed to be known, retrieval experiments are performed within-category. Street images may contain multiple garment items for retrieval. We consider each instance as a separate query for evaluation. Table 2 (left) shows the number of, query images, query items, shop images, and shop items.

Performance is measured in terms of *top-k accuracy*, the percentage of queries with at least one matching item retrieved within the first k results. Table 2 (right) presents the exact matching performance of our baselines and learned similarity approaches (before and after fine-tuning) for k=20. Whole image retrieval performs the worst on all categories. The object proposal method improves over whole image retrieval on all categories, especially on categories like eyewear, hats, and skirts, where localization in the shop images is quite useful. Skirts, for example, are often depicted on models or mannequins, making localization necessary for accurate item matching. We also trained category-specific detectors [10] to remove the noisy object proposals from shop images. Keeping the top 20 confident detections per image, we observe a small drop of 2.16% in top-20 item accuracy, while we are able to make the retrieval runtime up to almost an order of magnitude more efficient (e.g. 7.6x faster for a single skirt query on one core).

Our final learned similarity after category-specific fine-tuning achieves the best performance on almost all categories. The one exception is eyewear, for which the object proposal method achieves the best top-20 accuracy. The initial learned similarity measure before fine-tuning achieves improved performance on categories that it was trained on, but less improvement on the other categories.

	Source of distractors	
Category	Similar-to-query (%)	Similar-to-item (%)
Bags	77.3	81.6
Belts	65.5	53.9
Dresses	87.9	69.8
Eyewear	29.6	33.3
Footwear	58.9	44.1
Hats	69.8	57.0
Leggings	45.1	29.4
Outerwear	66.9	57.5
Pants	44.4	37.7
Skirts	69.4	66.6
Tops	78.1	66.1

Table 3: Human accuracy at choosing the correct item from different short-lists. Fig. 8 shows examples of the tasks.

Example retrieval results are shown in Fig. 6. The top three rows show success, where the exact matches are among the top results. Failure examples are shown in the bottom rows. Failures can happen for several reasons, such as visual distraction from textured backgrounds (e.g. 4^{th} row). A more accurate but perhaps more costly localization of the query item, might be helpful in these cases. Sometimes, items are visually too generic to find the exact item in shop images (e.g. blue jeans in the 5^{th} row). Finally, current deep representations may fail to capture some subtle visual differences between items (last row). We also observe errors due to challenging street item viewpoints.

Additionally, in Fig. 7 we plot the top-k retrieval accuracy over values of k for three example categories (dresses, outerwear and tops). For similarity learning, we vary k from 1 to the number of available items in the retrieved short list. For the baseline methods, we plot accuracy for k=1 to 50. We observe that the performance of our similarity network grows significantly faster than the baseline methods. This is particularly useful for real-world search applications, where users rarely look beyond the first few highly ranked results.

5.1. Human Evaluation

After developing automated techniques for finding clothing items from street photos, we then performed experiments to evaluate how difficult these tasks were for humans and to obtain a measure of how close the algorithms came to human ability. In these evaluations a human labeler was presented with the same query that would be given to an algorithm, and a set of possibly matching images. The task for the person was to select the correct item from the options. We use two criteria for determining what set of possibly matching images to display. Fig. 8 shows a query and two sets of possible shop photos.

As an initial measure of the difficulty of the task, we have people select a matching item for the query from the items in the dataset that are most similar to the *correct* item. We use whole image similarity to find those most similar items

Category	Queries	Query Items	Shop Images	Shop Items	Whole Im.	Sel. Search	Similarity.	F.T. Similarity
Bags	174	87	16,308	10,963	23.6	32.2	31.6	**37.4**
Belts	89	16	1,252	965	6.7	6.7	11.2	**13.5**
Dresses	3,292	1,112	169,733	67,606	22.2	25.5	36.7	**37.1**
Eyewear	138	15	1,595	1,284	10.1	**42.0**	27.5	35.5
Footwear	2,178	516	75,836	47,127	5.9	6.9	7.7	**9.6**
Hats	86	31	2,551	1,785	11.6	36.0	24.4	**38.4**
Leggings	517	94	8,219	4,160	14.5	17.2	15.9	**22.1**
Outerwear	666	168	34,695	17,878	9.3	13.8	18.9	**21.0**
Pants	130	42	7,640	5,669	14.6	21.5	28.5	**29.2**
Skirts	604	142	18,281	8,412	11.6	45.9	**54.6**	**54.6**
Tops	763	364	68,418	38,946	14.4	27.4	36.6	**38.1**

Table 2: Dataset statistics and top-20 item retrieval accuracy for the Exact-Street-to-Shop task. Last four columns report performance using whole-image features, selective search bounding boxes, and re-ranking with learned generic similarity or fine-tuned similarity.

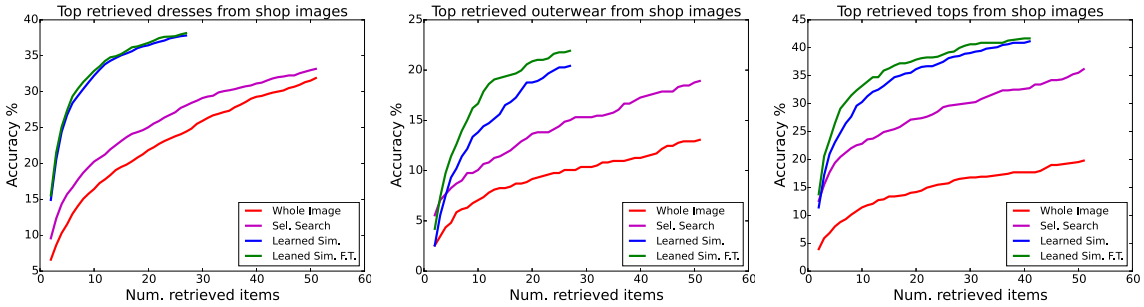


Figure 7: Top-k item retrieval accuracy for different numbers of retrieved items.

to the correct one. This set of possible choices is illustrated in the bottom part of Fig. 8. The human labelers select the correct item out of the 10 choices just over half of the time (54.2% averaged across all item types). This means that just under half the time, people do not pick the correct matching item, even out of a subset of only 10, albeit very similar, choices! This is one indication of the difficulty of the task. Table 3 shows results in the "Similar-to-item" column. The second human experiment is designed to temper our optimism about the success of the method. Here, we construct the 10 options to include the correct item as well as the 9 items most similar to the *query* according to our learned similarity, illustrated in the top part of Fig. 8. If the correct item was in the top 9, then we add the 10^{th}. Ideally, the images picked by our algorithm as good matches for the query, will be confusing to the human labelers and they will often pick one of these instead of the correct item. Alas, we find that there is still room for improvement. Consider dresses, where our algorithm does relatively well, picking the correct item in the top 10 in 33.5% of trials and getting the first item correct in 15.6%. In our human experiments, people pick the correct item out of 10 choices 87% of the time for dresses, which is significantly better. Table 3 shows results in the "Similar-to-query" column.

6. Conclusion

We presented a novel task, Exact Street to Shop, and introduced a new dataset. Using this dataset, we have evaluated three methods for street-to-shop retrieval, including our approach to learn similarity measures between the street and shop domains. Finally, we have performed quantitative and human evaluations of our results, showing good accuracy for this challenging retrieval task. These methods provide an initial step toward enabling accurate retrieval of clothing items from online retailers. Future work includes developing methods for more precise alignment between street and shop items for improving retrieval performance.

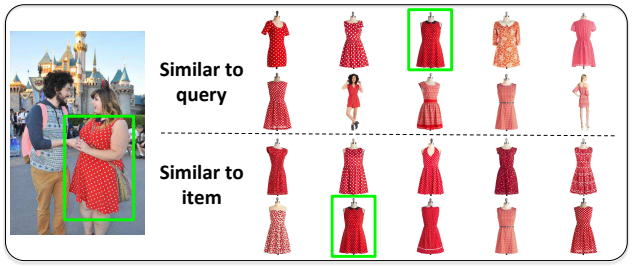


Figure 8: An example of our human evaluation tasks.

Acknowledgments. This work was partially supported by NSF IIS-1452851 and NSF IIS-1444234.

References

[1] S. Bell and B. Kavita. Learning visual similarity for product design with convolutional neural networks. In *ACM Transaction on Graphics (SIGGRAPH)*, 2015. 2

[2] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adap-

tation approach. *NIPS*, 2010. 2

[3] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. *ACCV*, 2012. 2

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*. 2012. 2

[5] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, , and N. Sundaresan. Style finder: Fine-grained clothing style recognition and retrieval. In *IWMV of CVPR*, 2013. 2

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014. 2

[7] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. *ICCV*, 2013. 2

[8] B. Fernando, M. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. *ICCV*, 2013. 2

[9] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *ACCV*, 2012. 2

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 7

[11] B. Gong, S. Yuan, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. *CVPR*, 2012. 2

[12] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. *ICCV*, 2011. 2

[13] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, 2015. 2, 5

[14] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell. One-shot adaptation of supervised deep convolutional models. *ICLR*, 2014. 2

[15] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. *CVPR*, 2010. 2

[16] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. ACM, 2013. 2

[17] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. *ECCV*, 2014. 2

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012. 5

[19] I. S. Kwak, A. C. Murillo, P. N. Belhumeur, D. Kriegman, and S. Belongie. From bikers to surfers: Visual recognition of urban tribes. *BMVC*, 2013. 2

[20] H. Lai, Y. Pan, Y. Liu, and S. Yan. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*, 2015. 2

[21] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. *ICCV*, 2013. 2

[22] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1), January 2014. 2

[23] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM MM*, 2012. 2

[24] 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 *CVPR*, 2012. 1, 2

[25] F. Perronnin and C. Dance. Fisher kenrels on visual vocabularies for image categorizaton. *CVPR*, 2006. 2

[26] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. *CVPR*, 2010. 2

[27] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. *ECCV*, 2010. 2

[28] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. *ECCV*, 2010. 2

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering facenet. In *CVPR*, 2015. 2

[30] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of beauty. *CVPR*, 2015. 2

[31] Z. Song, M. Wang, X.-S. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, 2011. 2

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2

[33] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 5

[34] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 2

[35] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, 2011. 2

[36] K. Yamaguchi, T. L. Berg, and L. E. Ortiz. Chic or social: Visual popularity analysis in online fashion networks. *ACM MM*, 2014. 2

[37] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: retrieving similar styles to parse clothing items. *ICCV*, 2013. 2

[38] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2

[39] J. Yosinkski, J. Clune, Y. Bengio, and H. Lipson. How transferrable are features in deep neural networks? *NIPS*, 2014. 2

[40] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015. 2, 5

[41] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, 2015. 2, 5