MovingFashion: a Benchmark for the Video-to-Shop Challenge

Marco Godi*[1], Christian Joppi*[1], Geri Skenderi*[1], and Marco Cristani[1,2]

[1]Department of Computer Science, University of Verona
[2]Humatics Srl, Verona, Italy
{marco.godi,christian.joppi,geri.skenderi}@univr.it
marco.cristani@{univr,humatics}.it

Abstract

Retrieving clothes that are worn in social media videos (Instagram, TikTok) is the latest frontier of e-fashion, referred to as "video-to-shop" in the computer vision literature. In this paper, we present MovingFashion, the first publicly available dataset to cope with this challenge. MovingFashion is composed of 14855 social videos, each one of them associated with e-commerce "shop" images where the corresponding clothing items are clearly portrayed. In addition, we present a novel baseline for this scenario, dubbed SEAM Match-RCNN. The model is trained by image-to-video domain adaptation, allowing the use of video sequences where only their association with a shop image is given, eliminating the need for millions of annotated bounding boxes. SEAM Match-RCNN builds an embedding, where an attention-based weighted sum of few frames (10) of a social video is enough to individuate the correct product within the first 5 retrieved items in a 14K+ shop element gallery with an accuracy of 80%. This provides the best performance on MovingFashion, comparing exhaustively against the related state-of-the-art approaches and alternative baselines[1].

1. Introduction

One of the most recent challenges in e-fashion is the so-called *video-to-shop* [1, 13], whose aim is to match a social video (Instagram, TikTok) containing one or more given clothing item(s), against an image gallery, potentially an e-commerce database (Fig. 2a,b). Individuating the outfit of a celebrity or social influencer can turn videos into priceless commercials, in a market where over a billion hours of video are uploaded and viewed on a daily basis [2].Video-to-shop derives from the *street-to-shop* problem, where the probe data is a single image [6]. On one hand, video-to-shop allows an increase of the available information by adding additional frames as probes. On the other hand, as shown in Fig. 2b, this information could be noisy due to challenging illumination, drastic zooming, human poses, missing data and multiple people (dis)appearing in the video. Another issue is that a video-to-shop system needs training data with millions of bounding box annotations, linking each box with a shop item [1, 13].

Our first contribution is MovingFashion, the very first publicly available video-to-shop dataset, composed by ∼15K different video sequences, each one related with at least one shop image. The videos of MovingFashion are obtained from the fashion e-shop Net-A-Porter (10132 videos) and the social media platforms Instagram and TikTok (4723 videos), and contain hundreds of frames per shop item, partitioned into a *Regular* and *Hard* setup.

Our second contribution is the SElf-Attention Multi-frame (SEAM) Match-RCNN, a video-to-shop baseline which individuates products and extracts features in a "street" video sequence by adopting a feature collection and aggregation mechanism, and then matching the products over a "shop" image gallery. SEAM Match-RCNN extends the popular Match-RCNN [4], state-of-the-art in the street-to-shop challenge, by applying image-to-video domain adaptation with the use of a novel Multi-frame Matching Head.

Technically, a pretraining on the image domain of the Match-RCNN enables it to provide initial pseudo-labels for a video sequence, individuating bounding boxes matching a particular product. The training on the target domain exploits our Multi-frame Matching Head, that aggregates features by means of a non-local block [11] between different frames, which in turn applies a temporal self-attention mechanism [3] and a scoring function. In this way an aggre-

*indicates equal contribution

[1] The code for SEAM Match-RCNN and the MovingFashion dataset are available here: https://github.com/HumaticsLAB/SEAM-Match-RCNN

Dataset	#Videos	#Traject.	#FramesXVideo [Avg.]	#Shops	[W, H]	#Pairs	Wild	Occlusion	Crowd	Available
AsymNet [1]	*526*	*26k*	***n.a.***	*85k*	***n.a.***	*39k*	***n.a.***	***n.a.***	***n.a.***	✗
DPRNet [13]	*818*	*5k*	***n.a.***	*21k*	***n.a.***	***n.a.***	***n.a.***	***n.a.***	***n.a.***	✗
MovingFashion	15k	15k	390	14k	[631± 12 , 770± 21]	15k	✓	✓	✓	✓

Table 1. Comparison of Video-2-shop datasets. *n.a.* stands for *not available*.

gation based on the attention score is used to create a single descriptor for a clothing item. In practice, SEAM Match-RCNN allows to train on video data where only the pairs <"street" video,"shop" image> are available, without annotated ground-truth bounding boxes. This policy permits to alleviate an intense annotation effort, which in the case of MovingFashion would have required drawing ∼18M bounding boxes. In the experiments, SEAM Match-RCNN gives the best performances on MovingFashion, against multiple baselines and state-of-the-art techniques. Actually, few frames (10) of a social video are enough to individuate the correct product within the first 5 retrieved items with an accuracy of 80%, making SEAM Match-RCNN a proof of concept for a potential product in e-fashion. Finally, SEAM Match-RCNN gives explainable results: thanks to self-attention visualization, we understand that the initial quarter of a social video carries the most information for guessing the correct product.

2. Related Work

Video-to-Shop. The first street-to-shop approaches employed single street images [4, 6, 9]; "street" video queries followed afterwards, paving the way for video-to-shop methods [1, 13]. AsymNet [1] aggregates frames by exploiting temporal continuity; it combines an LSTM and a binary tree, with each component requiring a separate training procedure. On the contrary, our SEAM Match-RCNN uses self-attention to learn a descriptor from a bunch of heterogeneous images, where temporal continuity is not required. DPRNet [13] manages the video-to-shop problem by treating it as street-to-shop, with a network that detects and tracks garments in the video, selecting automatically the frame with the highest quality (in terms of occlusions, blurring etc). That detection is finally fed into an image-to-image retrieval module. SEAM Match-RCNN does not perform this kind of tracking, which could be prohibitive on social videos that have strong heterogeneous variations on few frames.

Video-to-shop approaches shares similarities with video person Re-ID [8], where the goal is to match a video snippet of a person's silhouette against a gallery of image identities taken from a different camera. State-of-the-art approaches are VKD [10], NVAN [8] and MGH [12]. VKD proposes to learn using diverse views of the same target with a teacher-student framework, where the teacher educates a student who observes fewer views. NVAN is based on a non-local block self-attention module, embedded into the backbone CNN at multiple feature levels to incorporate both spatial and temporal characteristics of the pedestrian videos into the representation. Multi-Granular Hypergraph (MGH) is a novel graph-based framework which uses graph networks to cope with this problem.

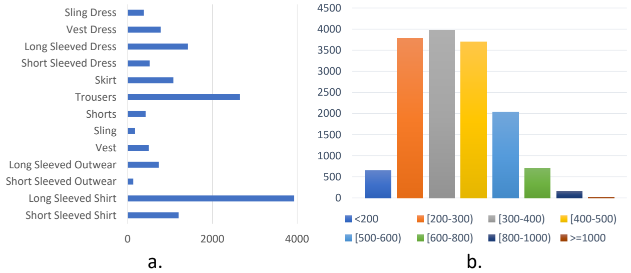

Figure 1. MovingFashion statistics; a) Cardinality of each clothing item class; b) Histogram of the number of frames for the street sequences.

Video-to-shop datasets. Unfortunately, no video-to-shop datasets are publicly available. The above quoted [1] and [13] use proprietary datasets, which have been not made open to the scientific community. We compare these datasets and their reported characteristics with our MovingFashion dataset in Table 1. It is visible that the datasets from AsymNet and DPRNet have a moderate number of sequences (526 and 818, respectively), while MovingFashion contains almost thirty times that amount (15K). In order to create more query data, DPRNet and AsymNet sample multiple sequences from the videos (generating 26K and 5K sub-trajectories, respectively). AsymNet contains 39K exact street-shop pairs and 85K diverse shop items, so one may infer shop distractors are present (shop items not present in the street set) but no details are provided on this. DPRNet has 21K Shop items, with no mentions about the exact pairs. MovingFashion has a single item associated with a unique shop image for each video, for a total of 15K unique (video) street-shop pairs.

The DeepFashion2 dataset (DF2) [4] presents a particular scenario: DF2 is made for the street-to-shop challenge, but some shop items are related to more than one street image (coming from different sources), creating 11K pairings. This provides us with another experimental setting.

3. MovingFashion

MovingFashion has 5.854M annotated frames, organized into 15045 video-shop *matching pairs*, i.e., each video is associated with a distinct *shop image*. In particular, there are 14.8K *unique* videos, among which some se-

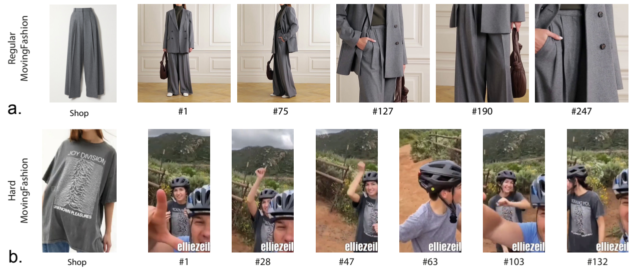

Figure 2. MovingFashion dataset samples. The top row contains a "Regular" sequence, the bottom row a "Hard" sequence.

quences (190 videos) have more than one associated shop item (*e.g.*, a t-shirt and trousers). The length of the videos is detailed in Fig. 1b, while the frame rate amounts to about 30FPS. Shop items are divided in classes, following the DeepFashion2 [4] taxonomy. The list of classes and the number of occurrences for each class in the dataset is reported in Fig. 1a.

3.1. Data sources

MovingFashion is formed by two subsets: *Regular* and *Hard*.

Regular MovingFashion: Regular MovingFashion consists of 10132 videos downloaded from the e-commerce website Net-A-Porter [2]: in the street video a single person is wearing the shop item in an indoor scenario (which can vary), and the corresponding shop image consists in the shop item captured over a plain background. This is the canonical shop image we have used in our experiments. Additionally, we have collected: a *front* shop image captured in the same background of the street video and worn by the same model in a frontal pose; a *rear* view image and a detail of the *fabric*. These last three were not used in the experiments. An example of Regular MovingFashion is showed in Fig. 2a (more in the supplementary material).

Hard MovingFashion: Hard MovingFashion consists of 4723 videos from the social platforms Instagram and TikTok. In this case, shop images have been obtained either by downloading images associated to the video as multiple images of the Instagram post or as part of the video itself (the spatial layout of some raw videos was organized in two halves, one being a still picture of the "shop" item, the other with the "street" video). Hard MovingFashion represents the hardest challenge, since all of the critical conditions listed in the introduction are present here, as also visible in Fig. 2b.

3.2. Data Collection and cleaning

All of the videos in Net-A-Porter have been designed to promote a clothing item, which made the data collection process simpler. Cleaning was necessary only to remove classes not compliant with the taxonomy of DeepFashion2 [4]. In contrast, Instagram and TikTok videos required a lot of work, starting with the search for the street videos and their shop counterparts using the available API, up to the careful scraping of hashtags and profiles. Other minor but time-consuming issues (fully/partially duplicate videos, wrongly associated shop items) are discussed in the supplementary material.

After collection and cleaning, we split the data into a training and a testing partition, taking care of applying the same split for each single class. We perform a 90/10 train/test split. Bounding boxes are extracted using a clothing detector. We then utilize the training data to train our SEAM Match-RCNN following the unsupervised procedure shown in Sec. 4.4. Since video sequences contain more than one item, to evaluate SEAM Match-RCNN and all the comparative approaches we create a tracklet containing the correct item for each street video sequence. That is done by selecting the tracklet that matches most with the shop item, following the *training tracking procedure* detailed in Sec. 4.4. A *tracklet* is a set of consecutive detections which refer to the same object. We manually check each one of these to ensure that at least 50% of the detections in the tracklet actually contain the shop item. For the detections which are too noisy (i.e. they do not focus on any precise clothing item), we dropped the entire sequence, in order to speed up the collection procedure. Fortunately, this happened on a minority of sequences ($\sim$150 videos), indicating a general success of the tracking procedure. The remaining tracklets are kept as noisy annotations in JSON format. All of the comparative approaches shown in Sec. 5 use these *ground truth tracklets* for training and testing.

4. SEAM Match-RCNN

SEAM Match-RCNN takes as input a sequence of street images $i_1 ... i_N$, and compares it with the gallery of K shop images providing a list of matching scores as output. Once the model has learned, the retrieval happens by means of three procedures: 1) *Tracklet creation*; 2) *Feature aggregation*; 3) *Video-to-shop matching*. Going through these steps will allow us to present the structure of the network, detailed in Fig. 3.

4.1. Tracklet creation

On the input video sequence we need to locate a set of consecutive detections which refer to the same object, dubbed here *tracklet*. Since multiple objects might be on the video, multiple tracklets are expected. The module that deals with this is the **Match-RCNN**, which is composed of three functions:

1. A clothing detector which provides convolutional fea-

[2] https://www.net-a-porter.com

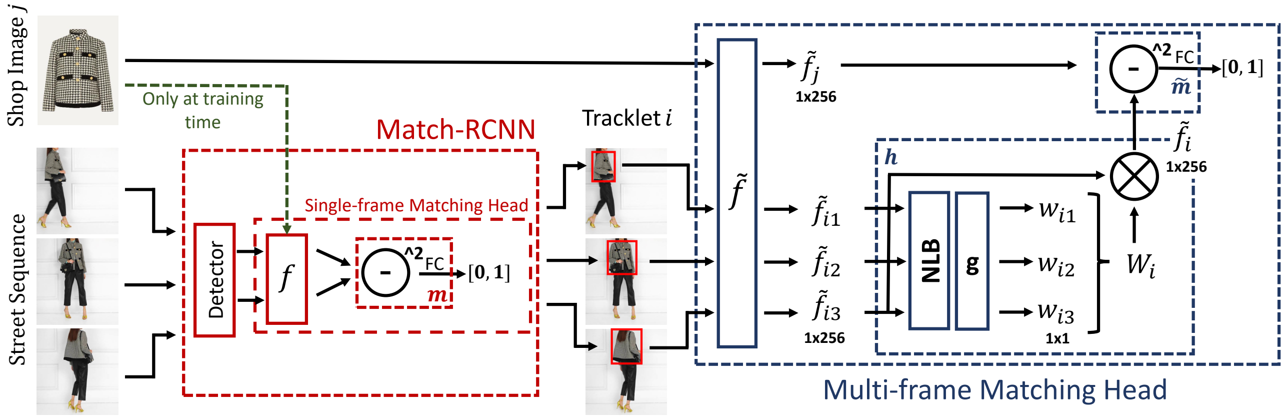

Figure 3. Architecture of our SEAM Match-RCNN system. Images are first processed by the Match-RCNN to extract bounding boxes and convolutional features. After tracking a clothing item across frames, its features are further processed by the Multi-frame Matching Head producing a final matching score between the street video sequence and the shop image.

tures $c_{i,t,k}$ with i indicating the i-th tracklet, t indicating the frame, k the k-th detection in that frame;

2. A 256-d feature extractor $f_{i,t,k} = f(c_{i,t,k}) \in \mathbb{R}^{256}$ which performs embedding of the convolutional features;
3. A matching score function $m(f_{i,t,k}, f_{i,t',k'}) \in [0, 1]$, comparing different embeddings.

f and m together form the **Single-frame Matching Head**.

The tracklet extraction procedure is performed in an iterative fashion, following a two-step process:

1. Determining the *pivot* bounding box: This represents the most confident detection $f_{i,t_{best},k_{best}}$ in the sequence and acts as the central reference based on which the tracklet will be built.
2. Performing *propagation* based on the *pivot*: By comparing the embedding of the pivot $f_{i,t_{best},k_{best}}$ with all of the detections in every frame, the tracklet i can be built. In particular, a detection joins the tracklet if its matching score (matching function m of the Single-Frame Matching Head) is above a certain threshold, to avoid considering frames where the item is not visible.

Once the tracklet i is built, its detections are removed, and another tracklet focusing on a different item can be built.

4.2. Feature aggregation

The next step is aggregating the information of a tracklet and condensing it into a single multi-frame descriptor. The module that deals with the feature aggregation procedure is the **Multi-frame Matching Head** and it is composed of the following functions and modules:

1. A 256-d feature extractor $\tilde{f}_{i,t} = \tilde{f}(c_{i,t}) \in \mathbb{R}^{256}$ operating on the bounding box at time t of the tracklet i, i.e., $c_{i,t}$.
2. A non-local block [11] module which applies self-attention, enriching $\{\tilde{f}_{i,t}\}_t$ with information coming from all the other bounding boxes related to the object tracklet i.
3. An attention module $g \colon \mathbb{R}^{N\times 256} \mapsto \mathbb{R}^N$ that for each detection in a tracklet computes an importance score w_t such that $\sum_t w_t = 1$.
4. An aggregation module, which fuses $\{\tilde{f}_{i,t}\}_t$ into a joint descriptor $\tilde{f}_i$ by a weighted average over the attention scores $\{w_t\}$: $h(x) = g(NLB(x)) \cdot x$, $x \in \mathbb{R}^{N\times 256}$.
5. A matching score function $\tilde{m}(\tilde{f}_j, \tilde{f}_i) \in [0, 1]$, which compares the aggregated descriptor for item k and street sequence i ($h(\{\tilde{f}_{i,t}\}_t)$ as $\tilde{f}_i$) with the the shop descriptor of image j.

The aggregation procedure starts with the feature extractor $\tilde{f}$, which creates the initial descriptors for each box in a sequence. Then, self-attention is computed by the non-local block module and afterwards the attention module calculates the attention weights for each descriptor. The aggregation module puts all of the above together, producing the single multi-frame descriptor $\tilde{f}_i$. Note that we discard temporal continuity *by design*. Social network videos usually have dramatic zooms, very fast pose dynamics and occlusions; moreover, elaborated videos may have shot changes which can fragment temporal continuity.

4.3. Video-to-shop matching

Following the feature aggregation procedure described above, we obtain a single multi-frame descriptor $\tilde{f}_i$ of the street tracklet i. In this final procedure, the matching score function $\tilde{m}$ of the Multi-frame Matching Head is used to match the aggregated multi-frame description with the single shop descriptor of image j, $\tilde{f}_j$ (which can be considered as a tracklet composed by a single frame), under the assumption that a single item is portrayed in the shop image.

We use the matching function $\tilde{m}$ on all the images in the shop gallery, producing in this way a list of matching scores between the street tracklet and all the images in the shop gallery, sorted in descending order, creating thus a *ranking*.

4.4. Model Training

To avoid the need of bounding boxes annotations, a time-consuming procedure especially for videos, SEAM Match-RCNN is trained by domain adaptation, through two phases: pretraining on the source image domain and training on the target video domain.

Pretraining on Source domain. The Match-RCNN part of SEAM Match-RCNN (detector and Single-frame Matching Head) is pretrained on an image street-to-shop dataset (e.g. DeepFashion2). The purpose of this phase is to train a model that is able to estimate bounding boxes and matching scores in the target domain (even with noisy predictions due to the domain gap). Such predictions are used to generate tracklets and pseudo-labels to train the Multi-frame Matching Head.

Training on Target domain. The training procedure estimates the parameters for the Multi-frame Matching Head of the SEAM Match-RCNN, whose structure is detailed in Sec. 4.2, and fine-tunes the Single-frame Matching Head, while the detector's weights are frozen. The weights of the features extractor $\tilde{f}$ and matching score function $\tilde{m}$ are initialized copying those of f and m from the pretrained Single-frame Matching Head. Conversely, the attention modules of h are randomly initialized. During training, image and street video sequence pairs (thanks to the MovingFashion pairing annotations) are sampled, which are leveraged in the tracking procedure (Sec. 4.1): the pivot selection is done by selecting the detection that matches the shop product the most in the whole video if the matching score inferred from the matching function m of the Single-frame Matching Head is over a certain threshold. The propagation step remains the same as in Sec. 4.1. With this *training tracking procedure* a tracklet is built such that, with a certain confidence, it contains the correct shop item due to the pivot selection starting from the ground truth shop image. This is considered as a positive match during training (i.e. we set 1 as a pseudo-label for the tracklet). For what concerns the Single-frame Matching Head fine-tuning, each detection that composes the tracklet is considered as positive match as well. The tracklet is then passed as input to the Multi-frame Matching head, which computes a singular multi-frame descriptor $\tilde{f}_i$ thanks to the aggregation procedure described in Sec. 4.2. In the end, this singular multi-frame descriptor $\tilde{f}_i$ is compared with the corresponding shop descriptor $\tilde{f}_j$ (obtained by using the feature extractor $\tilde{f}_j = f(c_j)$) utilizing the matching score function $\tilde{m}$. This produces a matching score in the range [0, 1].

Training is done by Stochastic Gradient Descent using cross-entropy loss for the classification of street videos and shop images as positive/negative matches. Positive pairings are built using the aforementioned procedure. All of the other combinations between tracklets and shop image descriptors are considered negative pairings (i.e. pseudo-label of 0) for the Single-frame Matching Head and the Multi-frame Matching Head.

5. Experiments

For the retrieval performance evaluation, we follow the testing protocol of DeepFashion2 [4] for evaluating a street image probe against a shop image gallery, with some modifications in order to cope with videos. In DeepFashion2, a street image generates multiple detections: each *street* detection can generate a *proper* matching with some shop image, if it overlaps by a threshold with the corresponding ground truth street bounding box and if its item class is correct, otherwise the matching score is 0.

On MovingFashion, we compute detections on every street image and we build tracklets using the *tracking procedure* detailed in Sec. 4.1. Then, we compute the average IoU between each street tracklet and the *ground truth tracklet*. The one with the highest average IoU is chosen and used as a query. In order to guarantee fairness in experiments, all baselines and comparative methods have been pretrained on two different street-to-shop datasets: DeepFashion2 and Exact Street2Shop [6]; the former has 873K probe-gallery pairs, while the latter 39K pairs only. Detailed results are reported for the first case, since performances were higher, while in the second case we show the main retrieval results, where our SEAM Match-RCNN remains the best performing approach.

5.1. Experiments on MovingFashion

We compare our technique with three types of approaches:

Multi-frame baselines. They are extensions of single-frame techniques to multi-frame. The *Max Confidence* [4] keeps the most confident detection in a tracklet and uses it for Single-frame Matching, employing a Match-RCNN. The *Max Matching* is inspired from [1] and considers the max matching score between the tracklet's street frames and each shop image. These two baselines are actually working with a single image, which is selected by looking at the entire pool of frames in a tracklet. They are also useful to validate the performance boost that comes when using multiple frames instead of single ones.

The *Average Distance* is inspired by [1] and consists in averaging single-image matching scores of the tracklet street frames and each shop image. The *SEAM Match-RCNN w/o NLB,g* is obtained by averaging *descriptors* (and not matching scores) together by average pooling, removing in practice the NLB self-attention block and the at-

Method	**MovingFashion**				**Regular-MovingFashion**				**Hard-MovingFashion**			
	T-1	**T-5**	**T-10**	**T-20**	**T-1**	**T-5**	**T-10**	**T-20**	**T-1**	**T-5**	**T-10**	**T-20**
Max Confidence [4]	0.29(*0.022*)	0.59(*0.020*)	0.72(*0.018*)	0.83(*0.017*)	0.31	0.63	0.76	0.86	0.21	0.46	0.60	0.71
Max Matching [1]	0.26(*0.025*)	0.60(*0.022*)	0.74(*0.021*)	0.85(*0.016*)	0.29	0.65	0.79	0.88	0.17	0.44	0.58	0.73
NVAN (2019) [8]	0.38(*0.023*)	0.62(*0.021*)	0.70(*0.022*)	0.80(*0.021*)	0.47	0.73	0.81	0.90	0.11	0.28	0.37	0.48
VKD (2020) [10]	0.40(*0.019*)	0.49(*0.019*)	0.58(*0.018*)	0.65(*0.019*)	0.49	0.59	0.68	0.75	0.13	0.20	0.27	0.34
MGH (2020) [12]	0.40(*0.021*)	0.59(*0.020*)	0.66(*0.020*)	0.74(*0.019*)	0.47	0.67	0.73	0.81	0.18	0.35	0.43	0.52
AsymNet (2017) [1]	0.42(*0.023*)	0.73(*0.019*)	0.86(*0.016*)	0.92(*0.011*)	0.49	0.81	0.93	0.96	0.22	0.47	0.65	0.74
AsymNet [AVG]	0.39(*0.023*)	0.66(*0.022*)	0.83(*0.019*)	0.90(*0.015*)	0.46	0.78	0.90	0.96	0.19	0.44	0.62	0.73
AsymNet [MAX]	0.40(*0.021*)	0.71(*0.020*)	0.81(*0.017*)	0.90(*0.014*)	0.47	0.80	0.91	0.95	0.20	0.42	0.61	0.73
Average Distance [1]	0.39(*0.021*)	0.73(*0.020*)	0.84(*0.017*)	0.91(*0.013*)	0.43	0.79	0.88	0.94	0.24	0.56	0.70	0.81
SEAM M-RCNN w/o NLB, g	0.37(*0.031*)	0.73(*0.025*)	0.86(*0.020*)	0.93(*0.015*)	0.42	0.78	0.90	0.95	0.21	0.57	0.75	0.85
SEAM M-RCNN w/o NLB	0.41(*0.021*)	0.73(*0.018*)	0.83(*0.015*)	0.91(*0.012*)	0.47	0.79	0.89	0.95	0.21	0.54	0.66	0.79
SEAM M-RCNN	**0.49(*0.022*)**	**0.80(*0.018*)**	**0.89(*0.016*)**	**0.94(*0.010*)**	**0.55**	**0.86**	**0.94**	**0.97**	**0.30**	**0.62**	**0.76**	**0.87**

Table 2. Video-to-Shop retrieval results on MovingFashion. Note: T-K means Top-K Accuracy.

Method	**T-1**	**T-5**	**T-10**	**T-20**
NVAN [8]	0.07	0.20	0.29	0.42
VKD [10]	0.16	0.24	0.31	0.38
MGH [12]	0.15	0.23	0.30	0.41
AsymNet [1]	0.09	0.26	0.37	0.49
SEAM Match-RCNN	**0.21**	**0.41**	**0.53**	**0.62**

Table 3. Top-K accuracy on MovingFashion, pretraining on S2S [6]

tention scoring function g from the SEAM Match-RCNN (see the scheme in Fig. 3). Finally, *SEAM Match-RCNN w/o* NLB keeps the attention score, without the self-attention. These last three are proper multi-frame baselines, in the sense that they merge information coming from multiple frames.

Video Re-Identification approaches. Video Re-Id approaches share similarities with Video-to-shop, in that they look for the best way to aggregate multi-frame information to match a person in a disjoint multi-camera setting. In practice, we consider each shop clothing item the equivalent of a Person Re-Identification Identity. The main differences between video-to-shop and Person Re-ID are that in our case less information is available in terms of pixels, since face and hair need to be discarded, focusing only on the clothing. Here we consider the SoA approaches of NVAN [8], VKD [10] and MGH [12][3].

Video-to-shop approaches. We consider the *AsymNet* [1] approach[4], and its modifications AsymNet[AVG] and AsymNet[MAX], in which the aggregations are made respectively by the average and the max of the similarity score nodes' outputs instead of using the fusion nodes binary tree. Asymnet exploits temporal continuity, yet it does not reach our results.

We set the sequence length to $T = 10$ for both training and testing, picking the frames using the Restricted Random Sampling strategy [7], thus ensuring coverage of the entire sequence length. To analyze variability in the results, we analyze the testing samples by sub-sampling them into pool of 800, 20 times, averaging the rankings and computing standard deviations.

Table 2 reports the results. Three facts become apparent: 1) As expected, single-frame approaches (Max Confidence, Max Matching) are definitely inferior ($<$15% on average) than multi-frame approaches; 2) The considered re-identification approaches, apart from top-1 scores, are inferior to genuine video-to-shop methods; 3) Our SEAM Match-RCNN surpasses all the competitors, including AsymNet, which gives a better aggregation than the AVG-distance in its [AVG] version and the MAX-distance in its [MAX] version. By looking at the ablative versions of SEAM Match-RCNN, one can note that the self-attention gives the strongest performance boost, followed by the attention layer. Their cooperation, i.e., the complete SEAM Match-RCNN, reaches the highest score.

By looking at the results within the Regular and Hard MovingFashion partitions, it is quite easy to note the general decrease in performance when it comes to the hard partition. To understand the performance qualitatively, Fig. 6 shows retrieval results from Regular (Fig. 6a) and Hard (Fig. 6b). Actually, even if Regular is apparently harder due to many shop alternatives which differ by fine grained results (see the flared jeans), the dramatic changes of poses and backgrounds of the Hard partition play a stronger role.

Failure cases arise when the original video has discriminant parts of the clothing item covered for most of the sequence, for instance the logo of the light blue sweatshirt (Fig. 4a). In this case, self-attention overlooks such important details. Complex visual patterns like the hard-rock band logo (Fig. 4b), seem to be not well characterized, meaning that the best match is attributed considering the shape of the logo rather than its content (the "Metallica" logo has the same shape of the probe logo).

The results w.r.t the single clothing classes of MovingFashion are reported in Table 4, where it is possible to observe our advantage in all but three classes. Interestingly, we found that the simpler the clothing in terms of

[3] At the moment of writing, the MGH approach is state-of-the-art in the MARS Video Person Re-Identification dataset, followed closely by VKD and NVAN.

[4] The code is available at `https://github.com/kyusbok/Video2ShopExactMatching`.

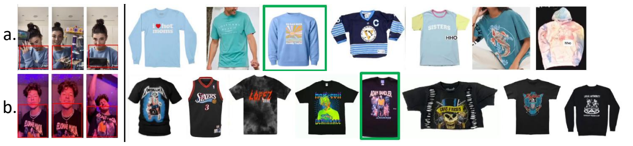

Figure 4. Failure cases results of SEAM Match-RCNN for the MovingFashion dataset. On the left, we show 3 frames sampled from the 10 frames used for aggregation. On the right the shop images retrieved with the corresponding rank. The correct matches are with a green border.

Categories	**NVAN** [8]	**VKD** [10]	**MGH** [12]	**AsymNet** [1]	**SEAM** **M-RCNN**
Short Sl. Shirt	**.46(.08)**	.20(.06)	.39(.06)	.35(.07)	.43(.08)
Long Sl. Shirt	.38(.04)	.44(.04)	.41(.04)	.45(.04)	**.44(.04)**
Short Sl. Outw.	.34(.20)	.23(.16)	.33(.18)	.35(.19)	**.42(.20)**
Long Sl. Outw.	.40(.10)	.43(.10)	.42(.10)	.36(.10)	**.46(.10)**
Vest	**.42(.12)**	.10(.07)	.24(.09)	.27(.11)	.31(.11)
Sling	.30(.19)	.16(.16)	.33(.18)	.32(.18)	**.36(.19)**
Shorts	.19(.13)	.27(.15)	.22(.13)	.25(.13)	**.39(.15)**
Trousers	.37(.05)	.28(.05)	.35(.05)	**.45(.06)**	.39(.05)
Skirt	.40(.08)	.52(.08)	.47(.08)	.39(.08)	**.56(.09)**
Short Sl. Dress	.34(.10)	.54(.11)	.35(.10)	.45(.11)	**.73(.10)**
Long Sl. Dress	.37(.07)	.63(.07)	.36(.07)	.57(.07)	**.68(.07)**
Vest Dress	.39(.09)	.49(.09)	.37(.09)	.42(.08)	**.64(.09)**
Sling Dress	.42(.14)	.39(.14)	.42(.13)	.32(.13)	**.69(.14)**
All Classes	.38(.02)	.40(.02)	.40(.02)	.42(.02)	**.49(.02)**

Table 4. Top-1 retrieval accuracy (and standard deviation) on MovingFashion for the 14 different item classes.

texture, the lower the retrieval performance. This is reasonable, since texture adds discriminative details, and this is why classes with simpler texture like vest, sling, shorts and trousers performed worse. We computed textureness by gray-level co-occurrence matrix contrast; quantitatively speaking, textureness and top-1 accuracy in retrieval are found to be correlated (Spearman $\phi = 0.72$, $p - value \leq 0.05$).

Another experiment regards the length of the sequences. Fig. 5 reports, with the associated error bars, the performance of SEAM Match-RCNN when increasing the number of frames from 1 to 20. As expected, the curves for both partitions, at both the top-1 and top-20 are increasing, with the "Hard" partition showing a plateau after 10 frames, while the "Regular" partition seem to benefit systematically. The reason could be that "Hard" sequences are dramatically noisy, and adding more frames will augment the clutter we need to deal with, while the "Regular" ones benefit because of the fine grained details which characterize the partition. Comparative performances when varying the sequence's length against other approaches are in Tab. 5. Notably, Asymnet [1] does not reach our results *even when doubling the number of input frames*.

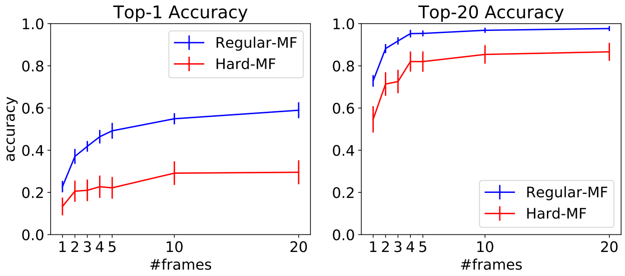

Figure 5. Plot of the SEAM Match-RCNN retrieval accuracy (y-axis) using different numbers of frames (x-axis) for aggregation. Error bars represent standard deviation of the accuracy.

Method	**5 Frames**	**10 Frames**	**20 Frames**
NVAN [8]	0.35	0.38	0.39
VKD [10]	0.36	0.40	0.43
MGH [12]	0.36	0.38	0.40
AsymNet [1]	0.37	0.42	0.44
SEAM Match-RCNN	**0.43**	**0.49**	**0.52**

Table 5. Top-1 accuracy on MovingFashion, with different number of frames.

Method	**T-1**	**T-5**	**T-10**	**T-20**
Max-Confidence	.19(*.014*)	.44(*.020*)	.54(*.020*)	.66(*.017*)
Max Matching [1]	.14(*.015*)	.45(*.020*)	.61(*.019*)	.75(*.017*)
NVAN (2019) [8]	.22(*.019*)	.37(*.019*)	.43(*.018*)	.49(*.019*)
VKD (2020) [10]	.21(*.014*)	.27(*.017*)	.33(*.017*)	.38(*.018*)
MGH (2020) [12]	.22(*.016*)	.34(*.018*)	.39(*.017*)	.45(*.019*)
AsymNet [GT] [1]	.21(*.016*)	.50(*.019*)	.62(*.017*)	.74(*.017*)
AsymNet (2017) [1]	.18(*.018*)	.43(*.020*)	.57(*.020*)	.70(*.018*)
AsymNet [AVG]	.16(*.016*)	.41(*.020*)	.54(*.020*)	.68(*.018*)
AsymNet [MAX]	.15(*.017*)	.42(*.020*)	.56(*.019*)	.70(*.017*)
Average Distance [1]	.22(*.017*)	.49(*.020*)	.63(*.019*)	.74(*.018*)
SEAM Match-RCNN w/o NLB, g	.20(*.016*)	.47(*.021*)	.60(*.021*)	.71(*.019*)
SEAM Match-RCNN [GT]	.30(*.013*)	.58(*.016*)	.67(*.016*)	.76(*.014*)
SEAM Match-RCNN	**.28(*.018*)**	**.54(*.020*)**	**.66(*.019*)**	**.76(*.017*)**

Table 6. Video-to-Shop retrieval results on MultiDeepFashion2. Note: T-K means Top-K Accuracy.

5.2. Experiments on unrelated sets of images

MovingFashion has street videos which depict clothing items in a variety of scenarios: indoor, outdoor, etc. We are interested in bringing this variety to the extreme, answering the following question: how does SEAM Match-RCNN behave when the street video sequence is formed by a few totally unrelated frames? In order to perform these experiments, we build Multi-DeepFashion2 from DeepFashion2 using the pairings between shop images and street sequences composed of multiple corresponding street images. The total pairings amount to 11K, each one composed of an image sequence (6 frames on average) sampled from different sources, along with the corresponding shop image. Results are in Tab. 6. Please note that, in order to be consistent with the 10-frames street sequence length we generate random repetitions for all the approaches given the smaller set of diverse images. The numbers indicate a decrease in general performance (less distinctive frames, more shop items); even in this case, we perform better than AsymNet.

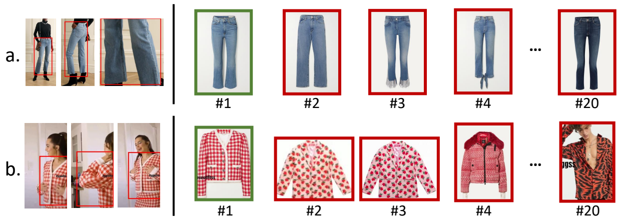

Figure 6. Qualitative retrieval results of SEAM Match-RCNN for the MovingFashion dataset. On the left, we show 3 frames sampled from the 10 frames used for aggregation. On the right the shop images retrieved with the corresponding rank. The correct matches are with a green border, otherwise red.

5.3. Experiments on the attention mechanism

The ablation studies of Table 2 clearly show that the attention layers play a crucial role for the SEAM Match-RCNN performance. Here we explain their role qualitatively and quantitatively. In Fig. 7 we report the attention values obtained after the application of the attention layer g to the output of the self-attention layer NLB of Sec. 4.2, i.e., $g(NLB(x))$. On row a), one can note that the attention is high when the heart logo is visible (0.31, 0.23 in the first two frames) and it goes down when it vanishes, despite the light blue shirt (last frame) being very similar area-wise. This means that the mechanism considers the heart logo as important for retrieval. On the second row b), the effect of an occlusion in the attention score (last frame). On the third row c), a white top with a logo gives a stable attention score (around 0.28). We manually cover the logo in the third frame, causing a clear decrease in the attention, uniformly increasing the ones highlighting the logo.

Finally, driven by best practices in social video editing [5], which state that a video message has to deliver its main content in the first 6 seconds to trigger the observers' attention, we calculate the attention every 5 percentiles on all the Moving fashion sequences, producing the curves in Fig. 8a) (on the whole MovingFashion dataset) and on the separate partitions Fig. 8b. Surprisingly, the data confirms this rule, showing a clear (Fig. 8a) peak around the first quartile (definitely within 6 seconds), then a decrease and a later increase with a local maximum on the fourth quartile. The same holds for the two separate partitions (Fig. 8b)), with less emphasis on the "Hard partition". The reason lies in the nature of the Net-A-Porter videos, which in many cases show the entire clothing item in the beginning of the sequence, with the model that moves subsequently, zooming up to critical detail (the belt for the shorts) towards the end (second peak). On the "Hard" partition, the attention for the clothing items is higher in the beginning, since the actors present their outfit and then exhibit their performance (dancing, gymnastics etc.), concluding in both the cases with uninteresting details clothing wise.

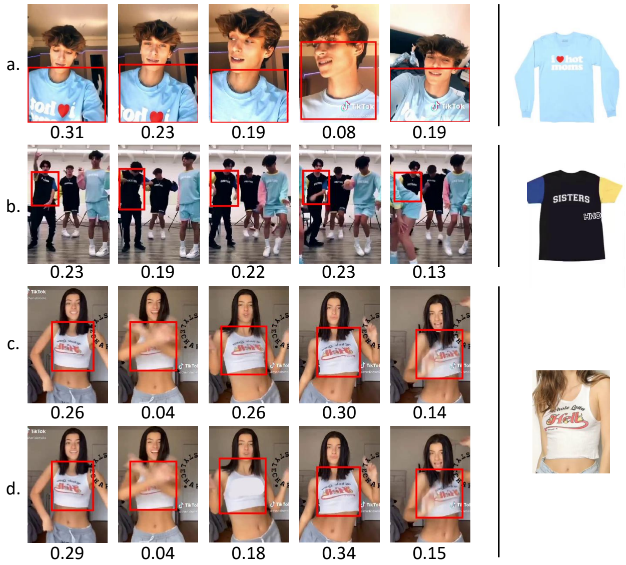

Figure 7. Qualitative observations on the attention behaviour. On the left, for each video sequence we show the detection bounding boxes and the computed attention score. On the right the paired shop item.

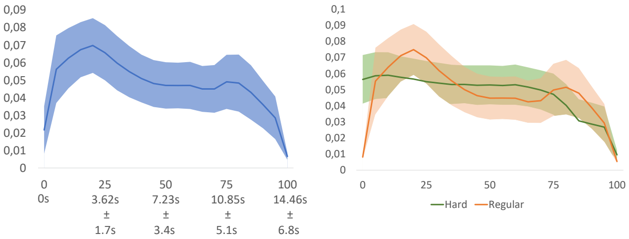

Figure 8. Mean attention score every 5 percentiles of the video length. For each video we sampled 21 equally spaced frames. On the left we report the average attention (y-axis) and frame-timing information (x-axis labels) for the whole MovingFashion dataset. On the right for the Regular and Hard subsets. We show error bands for the standard deviation.

6. Conclusions

Our SEAM Match-RCNN, trained on the new MovingFashion dataset, provides a strong baseline that shows video-to-shop matching can be performed on videos in the wild like TikToks, possibly unveiling fashion trends directly from social platforms and consequently attracting big fashion players.

Acknowledgments

This work is partially supported by the Italian MIUR through PRIN 2017 - Project Grant 20172BH297: I-MALL - improving the customer experience in stores by intelligent computer vision, and by the project of the Italian MIUR "Dipartimenti di Eccellenza 2018-2022". Thanks also to Giovanni Masotto for the collaboration in creating MovingFashion dataset.

References

[1] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4048–4056, 2017.

[2] Rodney Duffett. The youtube marketing communication effect on cognitive, affective and behavioural attitudes among generation z consumers. *Sustainability*, 12(12):5075, 2020.

[3] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018.

[4] 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.

[5] Maxwell Golling. Facebook video ads: Best practices for 2019, 2018.

[6] 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 *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.

[7] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[8] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *British Machine Vision Conference*, 2019.

[9] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[10] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *The European Conference on Computer Vision (ECCV)*, 2020.

[11] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[12] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Hongrui Zhao, Jin Yu, Yanan Li, Donghui Wang, Jie Liu, Hongxia Yang, and Fei Wu. Dress like an internet celebrity: Fashion retrieval in videos. In *proceedings of the International Joint Conferences on Artificial Intelligence*, pages 1054–1060, 07 2020.