Wardrobe Model for Long Term Re-identification and Appearance Prediction

Kyung Won Lee, Nishant Sankaran, Srirangaraj Setlur, Nils Napp, and Venu Govindaraju
Department of Computer Science and Engineering, University at Buffalo
{klee43, n6, setlur, nnapp, govind}@buffalo.edu

Abstract

Long-term surveillance applications often involve having to re-identify individuals over several days or weeks. The task is made even more challenging with the lack of sufficient visibility of the subjects faces. We address this problem by modeling the wardrobe of individuals using discriminative features and labels extracted from their clothing information from video sequences. In contrast to previous person re-id works, we exploit that people typically own a limited amount of clothing and that knowing a person's wardrobe can be used as a soft-biometric to distinguish identities. We a) present a new dataset consisting of more than 70,000 images recorded over 30 days of 25 identities; b) model clothing features using CNNs that minimize intra-garments variations while maximizing inter-garments differences; and c) build a reference wardrobe model that captures each persons set of clothes that can be used for re-id. We show that these models open new perspectives to long-term person re-id problem using clothing information.

I. Introduction

Person re-identification (re-ID) problem can be defined as finding a probe image of a person from gallery of images that have been taken either using different cameras [17], [23] or using a single camera at different time. As the number of surveillance cameras has increased, applications in video surveillance and human-computer interaction have become the main focus. At the same time, with the significant growth of e-commerce and on-line clothing shopping, clothing item retrieval has been receiving a great deal of attention in multimedia and computer vision communities.

Generally, person re-ID is difficult due to large appearance changes caused by environmental and geometric variations such as illumination, viewpoint, background, and occlusions. Utilizing detailed clothing information for modeling appearances of subjects is a novel means of performing re-ID, however it poses new challenges: 1) presence of ambiguous categories/attributes of clothes (e.g. hoodie vs jacket), 2) considering multiple layers of clothing (e.g. unbuttoned shirt over a t-shirt), 3) possibility that same clothes can belong to different persons (e.g. T-shirts with name-brand logos, basic solid color clothes, popular designs, etc.). Clothing has been used as a soft biometric for tracking and re-ID both implicitly [13], [2] and explicitly through color/texture appearance models [18]. However, this avenue has not been fully exploited. Clothing, specifically a wardrobe (collection of clothes), could be used to provide better re-ID performance over much longer time scales than a typical short-term tracking re-ID task where clothing items are assumed to be constant. By learning a wardrobe model for an individual, appearance in long-range camera views can be predicted. This approach complements other biometric modalities that require high quality images to function well. In this paper, we build an explicit wardrobe model and analyze its effectiveness as a soft biometric signal in long-term and long-range identification tasks. For this purpose, we

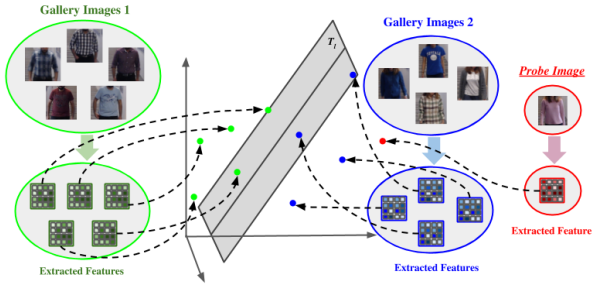


Fig. 1: Overview of Wardrobe Model: blue and green colors represent two different subjects and red color represents the probe frame. Circles represent the set of top and bottom clothing images of each subject. Colored dots represent attribute vectors from the Clothing Model. Hyperplane T_l depicts the classifier boundary.

978-1-5386-5541-2/18/$31.00

(i) collect data on which wardrobe models can be trained and tested; (ii) develop a reliable pipeline for person and clothing detection; (iii) develop a wardrobe model based on the detected clothing attributes; (iv) associate new data with previously built subject specific wardrobe models for the task of re-ID.

This paper is motivated by the main intuition that humans are able to largely identify people they know based on the clothing items worn by them. Especially in scenarios where the person's face is occluded, the appearance of the person, defined by the clothes and apparel they possess, provides important visual identification cues. We try to approach the re-ID problem in a similar way i.e. by modeling attire and their attributes. Specifically, we aim to assign semantically relevant labels (instead of relying on raw features) to the images of the subjects automatically which we then aggregate into a wardrobe model. Our inspiration for this approach is the on-line clothing retrieval system which searches the exact matched clothes for users using image search engines or shopping websites which has similar challenges as the person re-ID problem, viz., the diverse appearance of clothes, illumination, pose, cluttered background, viewpoint changes - from front, side or back, occlusion, different lighting conditions and common motion blur in videos. To the best of our knowledge, no attempts have been made to establish a benchmark for long-term re-ID by modeling wardrobe. We believe that this initial effort on person re-ID in long-term scenarios using purely information describing attire, could provide a novel and complementary approach to exiting methods and potentially provide new insights into long-term behavior by analyzing clothing choices. As a result, wardrobe modeling represents a significant contribution to tracking and surveillance based research and applications.

To demonstrate the approach, a longitudinal single camera tracking dataset is needed and hence, we prepared a new dataset addressing this requirement. Currently available person tracking / re-ID datasets are unsuitable for our approach since they are generally collected within a short duration of time (ranging from a few hours to a few days) with minimal repeated occurrences of individuals.

In summary, the contributions of this paper are two fold: (1) We collected a new fully annotated clothing attribute dataset, Indoor Long-term Re-identification Wardrobe (ILRW), for long-term person re-ID. (2) We address the problem of long-term person re-ID with attribute-based clothing labels as mid-level representations as it provides robust and semantic information than low-level descriptors.

II. Related Work

In this section we will briefly review the existing person re-ID techniques and also the different fashion datasets which are publicly available.

A. Person Re-ID

Two central components have been usually investigated for person re-ID: appearance modeling and distance metric learning. In appearance modeling, the design of visual features such as variations in color and texture histograms [5], [3], [20], Gabor features[11], [2], and others have been used. Also, local feature descriptors, such as SIFT [16], extracted from small sub-region patches in images, have produced good matching performance [24], [25]. Once these features have been extracted, different metric learning methods are generally applied. These methods learn Mahalonobis-like distance functions [7], [19], [1] which emphasize inter-personal distance and de-emphasize intra-person distance. The learned metric is used to decide whether a person has been re-identified or not. Also, other methods, such as learning decision tree ensembles [12] and salience [24], have been explored. Recently, Deep Neural Networks (DNN) have been successfully used for person re-ID [14], [13], [10]. [13] proposed an attribute-person recognition (APR) network which was built upon two baselines: one for person re-ID and the other for attribute recognition on two dataset namely, Market-1501 and DukeMTMC-re-ID. Also, [22] utilized wardrobe information for person re-ID. However none of these address the problem of long-term re-ID scenario. Moreover these methods don't restrict only to cloth based attribute with clothes-oriented descriptive labeling for re-ID.

B. Fashion Datasets

Fashion datasets are collected images from real-world consumer websites in fashion which are designed for clothing parsing and attribute estimation algorithms. Recently, there are many fashion datasets have been publicly available: Street2Shop [6], DARN (Dual Attribute-aware Ranking Network) [10], and DeepFashion [14], [15], etc. Also, there is a pedestrian attribute dataset [2] but it is specifically designed to use the attributes, such as gender, age, clothing style, in far-view surveillance scenarios. However, none of the works have explored long-term re-ID by using detailed-attribute levels of clothes.

III. ILRW Dataset Construction

We present a video dataset composed of 25 people, walking along an indoor corridor, with varied clothing appearances. In consideration for long term re-id we collected the data once every day, for 30 days, using a single camera. We term this dataset as "Indoor Long-term Re-identification Wardrobe (ILRW)" dataset.

In Figure 2, the left image depicts the position of the camera and the trajectory that each person takes and the

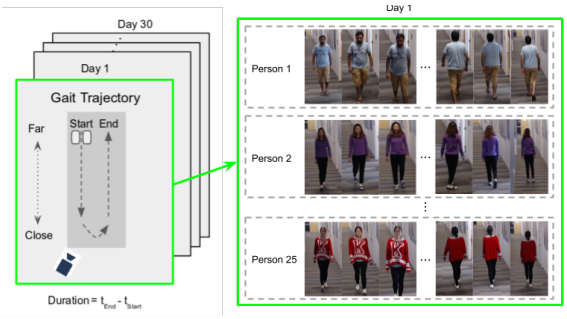


Fig. 2: Sample frames on one day from the ILRW dataset. Left image shows pedestrians' gait trajectory from the camera's perspective. The images on the right are the recorded samples of each pedestrian on a particular day.

TABLE I: Clothes Labeling Relation

Part	Attributes	Contents
Top	Category	t-shirt / sweater / hoodie / shirt / jacket
	Neck Style	v-neck / round-neck / collar / hood
	Fit	tight / medium / loose
	Formality	informal / semi-formal / formal
	Design	solid / vertical stripes / horizontal stripes / checked pattern / floral pattern / camouflage pattern / graphics / text / graphics with text / N-tone
	Length	short / long
	Sleeve Length	short / medium / long
	Reflection	yes / no
	Color	color histogram
Bottom	Category	jeans / skirt / sweat pants / shorts / long pants
	Design	solid / vertical stripes / horizontal stripes / checked pattern / floral pattern / camouflage pattern / graphics / text / graphics with text / N-tone
	Fit	tight / medium / loose
	Formality	informal / semi-formal / formal
	Color	color histogram

right image shows a few sample frames from 3 people captured on the first day. The subjects walk along a corridor with a web cam recording at 1920x1080 @ 30 fps from one corner. The resolution is sufficient to capture a person's appearance. Overall, we obtained around 3000 frames per person for 30 days, with 100 frames per day on average. After data collection, we apply the YOLO (You Only Look Once) [21] model to detect people from original images. We then use SSL (Self-supervised Structure-sensitive Learning) [4] for parsing clothing information to extract the top and bottom clothing sub-images. Finally, we manually annotate the image frames with the various clothing attributes as listed in Table I. Additionally, we provide the color histogram values for the sub-images.

IV. Proposed Model

We present two models: Clothing Models (CMs) and a Wardrobe Model (WM). CM predicts clothing attributes and WM manages every CM to identify people by utilizing clothing as biometric attributes.

A. Clothing Model (CM)

In order to generate automated labeling models for our Clothing Model (Fig. I), we use the segmented top and bottom images of the clothing produced by the SSL [4] framework. Separate DenseNet-121 Network [8] was trained to predict attributes of each category according to the clothing representation defined in Table I. DenseNet architecture was used since it only requires fewer parameters than traditional neural networks while maintaining comparable accuracy. In the DenseNet, a layer has access to all preceding features within its dense block, a block of directly connected layers. Direct connections from any layer to all subsequent layers are proposed in order to improve the information flow between layers. Mathematically, x_l the features produced by layer l are computed as $x_l = H_l([x_1, ..., x_l 1])$, where H_l denotes the operations of layer l and $[(\cdot)]$ represents the concatenation operator. The input images to these DenseNets were kept at 32x32 since it was giving notably well performing models. Also increasing the image size would require a higher computation power. By extracting the mid-level attributes of clothes from CM model,we are able to give a more semantic representation to the clothing apparels. This semantic representation provides a human interpretable aspect to clothing apparels, which is absent in other representations and additionally, it can also provide an advantage when performing clothing retrieval.

B. Wardrobe Model (WM)

The wardrobe of a subject is the collection of clothing items / apparel possessed by that individual. Observing the various combinations of clothes worn by an individual over time can help in constructing a robust model of that person's clothing choices and in turn, their proclivity for wearing certain types of clothing (eg. the tendency for a person to wear, say, darker colored t-shirts as opposed to other attire). Building such a wardrobe model could effectively aid person re-ID as an independent soft biometric.

We employ the clothing model to extract attribute labels for every frame of a day corresponding to an identity. To represent the subject's clothing attributes for the day, we perform a summarization operation over the individual frame-level clothing attribute predictions. The summarization essentially amounts to picking the most frequently occurring attribute value prediction over the set of video frames and hence acts as an efficient de-noising function over the inconsistent label predictions. The summarized attribute predictions are simply mapped to $[0, 1]$ by normalizing its range. In order to represent the color information contained in the frames, we computed the color histogram (in YUV space) using only the color from the pixels corresponding to the segmented region thereby

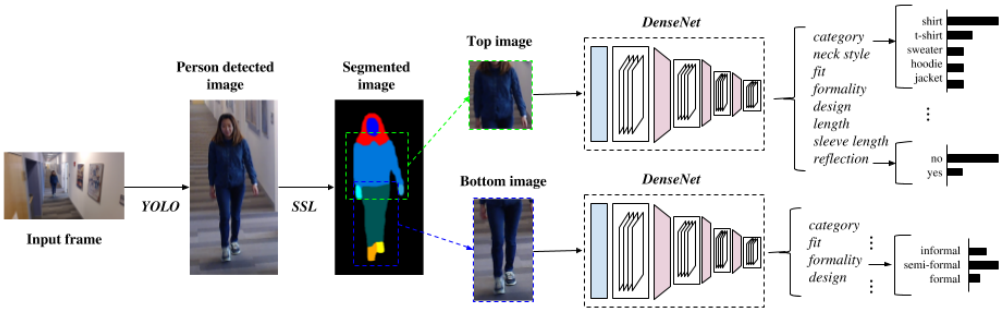


Fig. 3: Proposed Clothing Model(CM) framework for generating CM's candidate. Given a person detected image, DenseNet[9] predicts top/bottom categories/attributes from tightly bounded top/bottom images. By concatenating these values with color histogram, we could compose a feature vector which describes spatial and textural wardrobe information.

reducing the influence of background pixels. Additionally, the color histogram is normalized to sum to 1 which gives us a continuous probability distribution over the span of colors observed. To summarize the color information for the day, we perform a naive averaging of the computed histograms. These two components are concatenated to form the clothing representation for a day given a subject. The wardrobe model for the subjects are built by grouping their per-day clothing representations and given a probe image of top/bottom clothing item we use a multi-class linear SVM to identify the the wardrobe set it belongs to.

V. Experiments

Our experiments are conducted on the new ILRW dataset, since all the other pre-existing datasets do not serve the purpose of long-term Re-ID. We pre-processed each image by performing a histogram equalization and transforming it into the YUV color space. The data provided to the CM models are class balanced by augmenting the training data by using standard data augmentation techniques. Both the front and the back views of the clothing apparel are provided during training, so as to make the CM networks invariant of the view. All the CM networks were trained using mini-batch SGD with Nestrov momentum and a batch size of 32.

A. Analysis of the ILRW Dataset

Figure 4 shows various frequencies of observing the same clothes in N consecutive days under the following conditions: when a pedestrian wore the same (a) bottom clothes, (b) (inner and outer) top clothes, (d) only inner top clothes, and (e) only outer top clothes. In terms of top clothes, we examined inner and outer section individually and therefore could capture more semantic information of the days in the Fig. 4(d) and (e). (In Fig. 4(d), ambiguous represents an ambiguous situation when there's not enough data to observe inner top clothes in a frame.) Compared to the longest N days of top clothes in the Fig. 4(b),(d), and (e), the longest N days in the Fig. 4(a) shows a much higher frequency count because people change their bottom clothes less often than top clothes. Also, the longest N days in the Fig. 4(d) is larger than those in the Fig. 4(e). Among items within top clothes, this experiment shows that people typically have less number of outer clothes than inner clothes. Moreover, we observed that some people have a lesser variety of garments while other have a lot, especially bottom clothes compared to top clothes. Given that the ILRW dataset was collected in an unconstrained manner (subjects were free to follow their typical attire choices daily) this analysis highlights the need for longer durations of video tracking data to effectively be able to model people's wardrobe which is not available in the usual Re-ID datasets like DukeMTMC. In order to show the separability of the different clothing classes for top and bottom, Fig. 4(c) depicts confusion matrices of top and bottom. Since bottom clothes look more similar than top clothes, there are more interclass confusion in the bottom matrix. Fig. 4(f) also shows that as the number of days we observe subjects' clothing increases, there is an increase in the frequency of observing a previously seen clothing item, which again emphasizes the need for long term video tracking data.

B. Comparison of Clothing Models

We conduct comparisons using different Clothing Models (CMs) trained on two scenarios.

- Subject dis-joint training:- 10 random subjects where selected from the available 25 subjects. Clothing data of all the 30 days of these subjects was used to train the CM models. The testing of the models was performed on 30 days data of rest of the 15 subjects.
- Subject agnostic training:- 18 days data of *all* the subjects were used for training the CM models. The Testing is done on the rest 8 days data of all subjects.

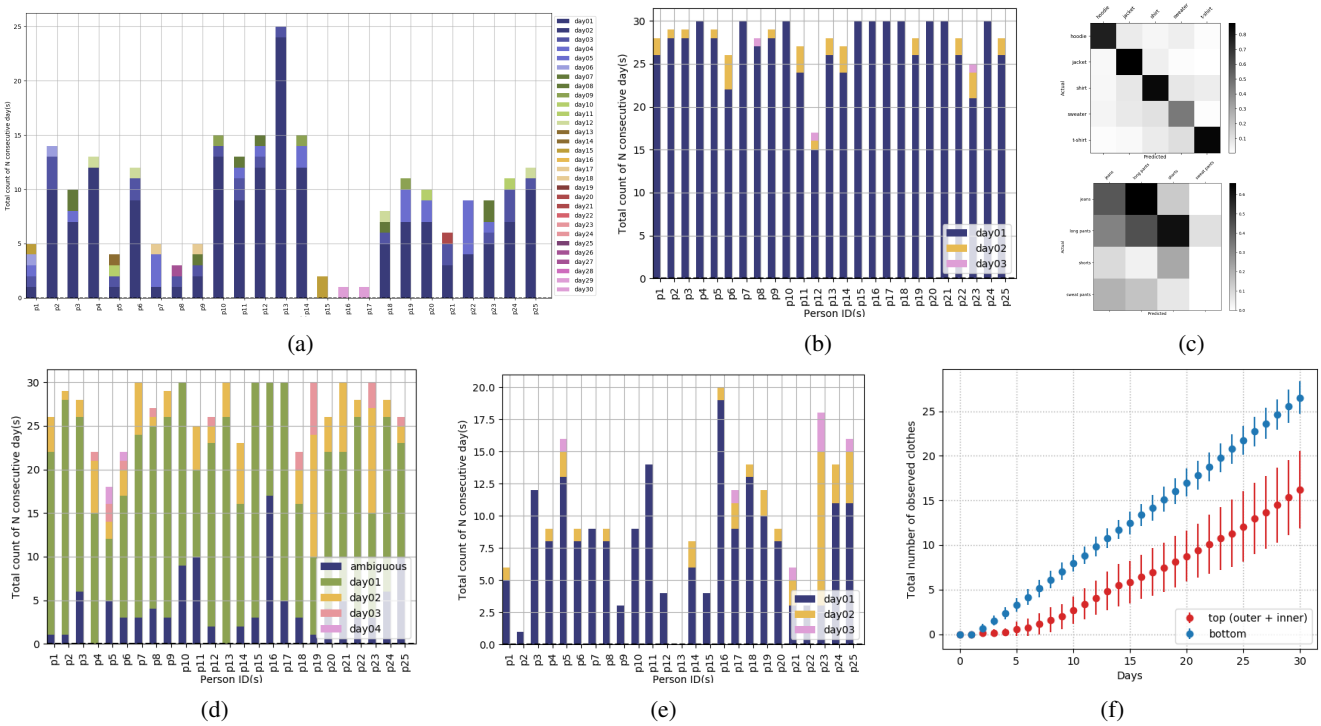


Fig. 4: (a) Total frequencies when a pedestrian wore the same bottom in N(1∼30) consecutive days, (b) total frequencies when a pedestrian wore the same top (inner and outer) in N(1∼3) consecutive days, (c) confusion matrices of top (up) and bottom (down) (d) total frequencies when a pedestrian wore the same top (inner) in N(1∼4) consecutive days including an ambiguous status, (e) total frequencies when a pedestrian wore the same top (outer)in N(1∼3) consecutive days, (f) total frequencies of pedestrians' clothes across 30 days

Parts		Top								Bottom			
Condition		Category	Fit	Neck Style	Formality	Design	Length	Sleeve Length	Reflection	Category	Design	Fit	Formality
S1	FBFC	55.36%	72.06%	57.65%	74.90%	60.37%	95.48%	86.47%	95.61%	51.55%	75.53%	70.61%	69.81%
	SC	**69.78%**	**77.78%**	**83.11%**	**83.11%**	**75.33%**	**96.44%**	**91.33%**	**96.44%**	**60.44%**	**80.00%**	**80.44%**	**82.22%**
S2	FBFC	80.29%	85.12%	78.96%	86.12%	81.40%	98.70%	92.35%	98.44%	76.99%	95.25%	84.07%	89.48%
	SC	**87.54%**	**92.26%**	**93.27%**	**91.58%**	**90.91%**	**99.66%**	**94.28%**	**99.66%**	**81.82%**	**98.32%**	**88.55%**	**93.60%**

TABLE II: Accuracy of CM on the first scenario (S1) and the second scenario (S2)

Table II (S1) shows the results of subject dis-joint scenario and Table II (s2) shows the results of the subject agnostic scenario. The performance of most of the CM models drops in subject disjoint scenario compared to to the subject agnostic scenario. This can be attributed to the limited prediction capability of the models, when presented with a completely unseen clothing apparels.

The CM models evaluation were done using two methods a) Frame by frame comparison(FBFC) in which prediction of the CM models are done at each frame of the test data and the accuracy is reported. b) Summarized comparison(SC) in which the prediction of CM models are done for each frame and then aggregated so as to get a unified representation. This representation is compared to the ground truth and accuracy is reported. It can be observed from the Table II that summarized comparison (SC) gives an better overall performance in predicting clothing attributes.

C.Wardrobe model based re-ID

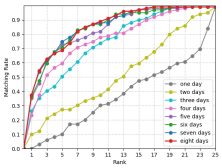
In Figure 5(b), we can observe that as the total number of previously detected clothes increases as accumulated days of being trained becomes larger, the results for person re-ID increases proportionately. This has demonstrated that by using an effective clothing model and building a wardrobe model off of its predictions, we can use simple classification techniques such as SVMs to produce appreciable re-ID performance.

VI. Conclusion

This paper provides additional avenues for re-ID in long-term scenarios by employing a semantic appearance representation based on people's clothing. For long-term appearance prediction, the Indoor Long-term Re-

Rank	Accuracy
1	0.37
3	0.63
5	0.75
10	0.88

(a)



(b)

Fig. 5: Wardrobe model based re-ID performance: (a) Rank level re-ID performance using the proposed model (b) Rank level re-ID performance for varying number of days observed to build model.

identification Wardrobe (ILRW) has been collected and comprehensively annotated. ILRW contains over 70,000 images with 12 semantic attributes for top and bottom as mid-level representations. Two major models, Clothing Models(CM) and Wardrobe Model(WM), are built for the analysis of recognizing and retrieving clothes in the person re-ID domain. We conclude that wardrobe models can be built effectively by using observations over several weeks, and that once they are built can be used for identification. This dataset is limited in the type of participants, and the required timescale for building wardrobe models might differ between populations. The proposed method's efficacy on the ILWR dataset motivates wider studies in a more general setting. For future work, more complex clothing models that do not assume at most 2 layers might be useful, as well as using multi-camera observations to more reliably build clothing models. A systematic study about the impact of low-quality images and less constrained environment/scenario would provide insights into the potential for real-world deployments. Larger datasets would also allow studying identifiability of clothing, e.g. what fraction of clothing items or combinations are distinctive. The wardrobe modeling approach also opens the door for entirely new tools in long-term, long-range surveillance applications. For example, analyzing clothing choices people make *given* their wardrobe, or the composition of the wardrobe itself, might provide insight into a person's state of mind or expose other behavioral or social information.

VII. Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant IIP #1266183.

References

[1] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *12th European Conference on Computer Vision*, volume 7574, pages 806–820. Springer, 2012.

[2] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM, 2014.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[4] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and A new benchmark for human parsing. *CoRR*, abs/1703.05446, 2017.

[5] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[6] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[7] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, pages 780–793. Springer, 2012.

[8] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[9] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3, 2017.

[10] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. *CoRR*, abs/1505.07922, 2015.

[11] W. Li and X. Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[12] Y. Li, Z. Wu, and R. J. Radke. Multi-shot re-identification with random-projection-based random forests. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 373–380. IEEE, 2015.

[13] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[14] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[15] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision (ECCV)*, 2016.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[17] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE, 2009.

[18] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.

[19] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR, 2012 IEEE Conference on*, pages 2666–2672. IEEE, 2012.

[20] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[21] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[22] Y. Takahashi and H. Miyano. A compact color descriptor for person re-identification with clothing selection from a wardrobe. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[23] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

[24] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR, 2013 IEEE Conference on*, pages 3586–3593. IEEE, 2013.

[25] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151, 2014.