DEEP LEARNING PROTOTYPE DOMAINS FOR PERSON RE-IDENTIFICATION

Arne Schumann[⋆] *Shaogang Gong*[†] *Tobias Schuchert*[⋆]

[⋆] Fraunhofer IOSB, Karlsruhe, Germany
[†] Queen Mary University of London, UK
{arne.schumann,tobias.schuchert}@iosb.fraunhofer.de, s.gong@qmul.ac.uk

ABSTRACT

Person re-identification (re-id) is the task of matching multiple occurrences of the same person from different cameras, poses, lighting conditions, and a multitude of other factors which alter the visual appearance. Typically, this is achieved by learning either optimal features or distance metrics which are adapted to specific pairs of camera views dictated by the pairwise labelled training datasets. In this work, we formulate a deep learning based novel approach to automatic *prototype-domain* discovery for domain perceptive person re-id. The approach scales to new and unseen scenes without requiring new training data. We learn a separate re-id model for each of the discovered prototype-domains and during model deployment, use the person probe image to *automatically* select the model of the closest prototype-domain. Our approach requires neither supervised nor unsupervised transfer learning, i.e. no data available from target domains. Extensive evaluations are carried out using automatically detected bounding boxes with low-resolution and partial occlusion on two large scale re-id benchmarks, CUHK-SYSU and PRW. Our approach outperforms state-of-the-art unsupervised methods significantly and is competitive against supervised methods which use labelled test domain data.

***Index Terms*—** Person Re-Identification, Deep Learning, Transfer Learning, Prototype Domain

1. INTRODUCTION

The task of re-identifying the same person across different cameras has attracted much interest in recent years. Most existing approaches interpret each camera as a separate visual domain and focus on developing features or metrics that can robustly recognize a person across such *camera-view-perspective* domains [1]. In this work, we consider other *camera-view-independent* factors, such as pose, illumination, occlusions, and background which influence the visual appearance of a person. We aim to identify visual domains defined by these factors and use them to construct *camera-view independent re-id* models for better scalability to unknown camera views and scenes.

We propose a two-stage approach to automatically discover visual domains in large amounts of diverse data and use them to learn feature embeddings for person re-identification. In the first stage, we create a training dataset with a large degree of visual variation by pooling many existing re-id datasets. We then apply clustering based on feature learning in convolutional neural networks (CNNs) to automatically discover dominant visual domains (prototype domains). In the second stage, we apply CNNs to learn a feature embedding for each of these prototype domains. This allows our approach to learn specific details about each individual prototype domain while ignoring the complexities of others. For example, an embedding learned for a domain which predominantly contains people of dark-dress does not need to encode information relevant to distinguishing a person dressed in light blue colors from a person dressed in white clothes. The domain perceptive embedding can thus focus on learning more subtle discriminative characteristics among similar visual appearances. At test time, a probe image is first matched to its closest domain. Then, the feature embedding learned on that domain is used to perform re-identification. Note, this approach is purely *inductive*. It does not require any training data (labelled or unlabelled) from the target (test) domains, and the model is designed to scale to any new target domain. Our approach is particularly well suited to scenarios in which no fixed set of camera views is available (*i.e.* no fixed domain borders are specified). We thus evaluate it on the latest CUHK-SYSU and PRW datasets, which contain images from diverse sources of mobile cameras, movies and fixed-view cameras, with a multitude of view angles, backgrounds, resolutions and poses. Our inductive approach yields state-of-the-art accuracy on CUHK-SYSU and performs very close to state-of-the-art *supervised* approaches on PRW.

Our contributions are: **(1)** We formulate a novel approach to automatic discovery of prototype-domains for charactering a person's visual appearance with domain perceptive awareness. **(2)** We develop a deep learning model for domain perceptive (DLDP) selection and re-id matching in a single automatic process without any supervised or unsupervised domain transfer learning. **(3)** We show the significant advantage of our model by outperforming the state-of-the-art on the CUHK-SYSU benchmark [2] with up to 5.6% at Rank-

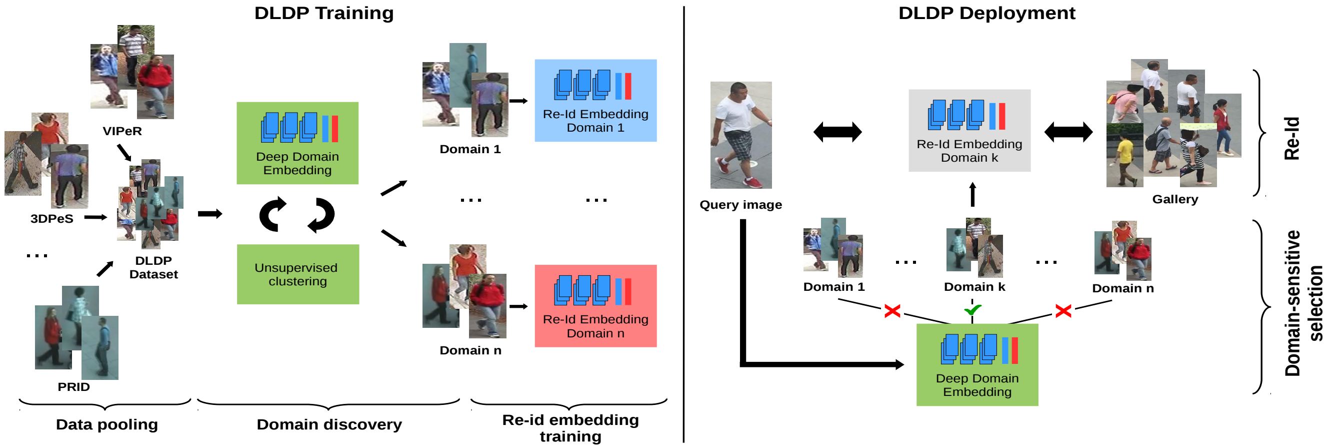

Fig. 1. DLDP overview: For training we discover prototype-domains in diverse data and learn a domain-specific re-id model for each domain. At deployment, the query image identifies a matching domain and optimal model for re-id.

1 re-id, and being competitive on the PRW benchmark [3] of 45.4% Rank-1 re-id compared to the 47.7% state-of-the-art, notwithstanding that the latter benefited from model learning on target domain.

2. RELATED WORK

Most re-id approaches fall in two categories: Feature design and metric learning. The former aims to develop a robust feature representation. The latter focuses on optimizing a distance metric that, given any feature, yields small distances for matching person images and large distances for images of different people. In recent years, deep learning has shown superior performance for image classification, and has been applied to person re-id as well.

Many deep learning approaches focus on feature learning. CNNs are trained using either the softmax [4, 5, 6] loss for classification of person IDs, ranking losses, such as the triplet loss [7], or modifications thereof [8]. The responses of specific network layers are then used as features (feature embeddings) for re-id. Some deep learning approaches focus on studying network layers specifically designed for person re-id, such as the filter pairing approach by Li *et al.* [9] or the neighborhood matching layer [10, 11]. Xiao *et al.* [2] propose an approach which combines person detection and re-id into an end-to-end model.

A few works have also focused on how to better separate individuals with very similar visual appearance. Karaman *et al.* [12] combine simple discriminants with a Markov Random Field which leverages local structure in feature space. Garcia *et al.* [13] introduce a re-ranking method which uses content and context clues to optimize visually similar top ranks.

Recent studies have addressed cross-domain re-id by using target domain data for supervised [14, 15, 16] or unsupervised [17, 18] domain adaptation. Others have evaluated their models across datasets without any adaptation to the target domain [19, 20, 4]. To our knowledge, the proposed model in this work, uniquely, does not rely on domain adaptation using target domain data whilst learning domain perceptive re-id for unknown target domains.

3. METHODOLOGY

The central objective of our approach is to learn a domain adaptive re-id model (domain perceptive) which is scalable to new and unseen domains without requiring any additional training data for adaptation. We propose a two-stage approach to achieve this: (1) In the first stage, characteristic and dominant prototype domains are automatically discovered in large amounts of diverse data; (2) In the second stage, this information is used to train a number of domain specific re-id embeddings by deep learning. An overview of our approach is given in Figure 1.

3.1. Automatic Domain Discovery

Divergent Data Sampling: A key requirement for a meaningful domain discovery is *divergent data sampling* which aims to provide a large range of realistic visual variation. In order to achieve such a high degree of variation, we pool a number of publicly available person re-identification datasets into a new, large dataset for domain discovery, called DLDP dataset[1]. We combine 10 datasets: HDA [21], GRID [22], 3DPeS [23], CAVIAR4REID [24], i-LIDS [25], PRID [26], VIPeR [27], SARC3D [28], CUHK2 [29], and CUHK3 [9]. The combined dataset contains images of 4,786 different persons with a total of 41,380 bounding boxes of which 14,097 are obtained by a person detector. We resize all bounding boxes to a uniform size of 160 × 80 pixels.

Network Architecture: We use the same network architecture for all models trained in our approach. The architecture is based on inception layers [30] and resembles that described

[1] The DLDP dataset will be made publically available.

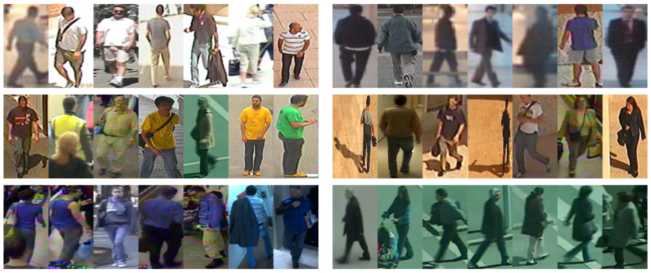

Fig. 2. Prototype-domains discovered by our approach: Supervised initialization with a re-id net (left), initialization by weights learned through unsupervised autoencoding (right). Our model learns semantically meaningful prototype-domains (*e.g*. light-colored, yellow and blue clothing).

in [5]. We empirically found two modifications to this architecture which result in increased accuracy for our approach: we add a fourth initial convolutional layer and increase the size of the final feature layer to 512.

Prototype-Domain Discovery: We apply deep learning based clustering to discover prototype domains from the DLDP dataset. Specifically, we exploit the concept of unsupervised deep embedding space learning proposed in [31]. Our deep learning clustering model alternates between (1) training a CNN to learn a feature embedding and (2) applying conventional k-means clustering in the embedding space to find clusters. To initialize the weights of our feature embedding CNN, we first train the model using the person ID labels available in the data. We set the last fully connected layer of the network to 4,786 dimensions and train using person ID labels in a one-hot encoding and a softmax loss for ID classification. The resulting weights are used for initialization of the domain embedding. This initialization from a re-id pretrained net is crucial to the success of our prototype domain discovery. The re-id training ensures that the initial model does *not* react strongly to the dataset biases present in our feature pool. *This prevents the clustering from simply discovering trivial dataset boundaries as prototype domain boundaries* and instead, lets the model focus more on the content of each person bounding box. Figure 2 shows three domains resulting from our proposed initialization compared to three domains resulting after initialization from an autoencoder. For the domain discovery (*i.e*. k-means clustering) the ID softmax loss layer is replaced by a softmax loss which corresponds to the number of clusters. Thus, after a supervised initialization, the domain discovery continues in an unsupervised manner.

Training Strategy: For training our deep clustering model we use a low initial learning rate of 1e-3 to ensure that the cluster embedding does not deviate too quickly from its initialization. Given the initial embedding, we perform 25 repetitions of k-means clustering in the embedding space and select the best result for the next refinement of the embedding. This ensures stability of the iterative training process. The refinement (fine-tuning) of the embedding CNN is then performed for a further 10,000 training iterations. We divide the learning rate by 10 every two iterations of the discovery process. This iterative process is repeated until less than 1% of images change their cluster assignments.

3.2. Domain Perceptive Re-Id Model

The second stage of our DLDP re-id model consists of learning a domain-sensitive re-id model for each prototype domain. That is, we train one feature embedding *with all person ID labels* for each of the discovered prototype domains from the first stage. To that end, we start by training a common generic baseline re-id model on all available data without considering the domains. The individual domain models are then trained by fine-tuning this baseline model.

Baseline Model: For a baseline, we train a model for 60,000 iterations on all training data to learn a generic feature embedding without domain specific adaptation. The initial learning rate is set to 1e-2 and divided by 10 after every 20,000 iterations. We use the output of the 512 dimensional layer as our baseline feature embedding for re-id. Re-id matching is performed by cosine distance.

Domain Sensitive Embeddings: In order to create domain specific training data, we select for each person ID in a given domain all of that person's images and add them to the training data for the domain. This data sampling method allows the domain models to specialize and focus particularly on the visual cues relevant to persons from their domain while not having to also learn how to distinguish between persons from other domains. We train the domain specific embedding by fine-tuning the baseline model. The dimension of the softmax layers is adapted according to the number of persons in each domain. For each domain we fine-tune for 30,000 iterations at an inital learning rate of 1e-4.

Automatic Domain Selection: During model deployment, a probe person image is first matched to its most likely domain by the deep clustering model (Section 3.1). The corresponding domain specific re-id model is then used to rank the gallery images. This dynamic model selection at query time lets our model adapt to the target domain on-the-fly without any need for re-training.

4. EXPERIMENTS

Datasets and Protocol: We evaluate our model on the CUHK-SYSU [2] and PRW [3] full-image datasets which consist of 8432 and 932 person IDs as well as 99,809 and 34,304 bounding boxes, respectively. Both datasets contain a large number of viewing angles, range of pose, occlusions and resolution. This allows us to investigate the generalization capability of our approach under very realistic conditions, its ability to handle large amounts of varying views and to evaluate its performance on automatically detected person bounding boxes. Note, that both datasets contain many distractor persons without ID in the galleries (*i.e*. *open-set* eval-

	k=1	k=2	k=4	k=6	k=8
mAP	68.4	67.1	71.4	72.6	74.0
Rank-1	70.3	68.7	73.3	75.1	76.7

Table 1. Effect of prototype-domain numbers (k) on re-id rate, using CUHK-SYSU with the gallery 100 setting.

uation) and any person detector will generate false positive detections which the re-id approach has to handle. Since our approach does not require training data on the target domain, we only use the test part of each dataset. We follow the exact evaluation protocols specified in [2] and [3], respectively and use the provided evaluation code. We use mean Averaged Precision (mAP) and Rank-1 accuracy as evaluation metrics for comparison to existing models [2, 3].

Number of Domains: In Table 1 we investigate the influence of the number of chosen domains on the accuracy of DLDP. The setting for a single domain (k=1) corresponds to our baseline model. For few domains (k=2) the resulting re-id models perform less accurate than the baseline. This is due to the low degree of specialization in the domains which leads to the resulting models merely being weaker versions of the baseline model. Given an increasing number of domains, DLDP's advantage becomes greater until it saturates around eight domains, which is adopted for all other experiments.

		mAP	Rank-1
[2]	Person Search [2]	55.7	62.7
	Baseline Model	61.4	68.3
	DLDP	**66.8**	**71.9**
SSD	DLDP (SSD VOC300)	49.5	57.5
	DLDP (SSD VOC500)	**57.8**	**64.6**
	Baseline Model (SSD VOC500)	54.2	59.9

Table 2. Performance comparison between DLDP and [2] using auto-detections from [2] or the SSD detector.

		mAP	Rank-1
DPM Inria	IDE [3]	13.7	38.0
	IDE$_{det}$ [3]	**18.8**	**47.7**
	BoW + XQDA [3]	12.1	36.2
	Baseline Model	12.9	36.5
	DLDP	15.9	45.4
SSD	BoW + XQDA (SSD VOC300)	6.8	26.6
	DLDP (SSD VOC300)	10.1	35.3
	DLDP (SSD VOC500)	**11.8**	**37.8**

Table 3. Performance comparison on the PRW dataset, with 5 bounding boxes per image. Note that all existing models except ours were trained (supervised) on the PRW dataset.

Comparison with the state-of-the-art: We give results on the CUHK-SYSU dataset for gallery sizes of 100 images in Table 2. Our baseline model outperforms [2] when relying on the same set of detections and still performs very well when relying on detections generated by the SSD detector [32]. The full DLDP approach outperforms [2] in both settings. Importantly, DLDP in combination with the SSD detector (trained on Pascal VOC data) still outperforms [2] (which relies on re-id features *and* detections which were trained on the CUHK-SYSU training set) by 1.9% at Rank-1.

For the evaluation on the PRW benchmark, we compared DLDP to a baseline using BoW features and XQDA metric learning [33] and two deep feature embeddings IDE and IDE$_{det}$ from [3] which are based on the AlexNet [34] architecture, trained on ImageNet and fine-tuned for re-id on PRW. For person detection, we used both the DPM person detector [35] trained on the INRIA dataset [36] provided by [3] and the SSD detectors for a fair comparison. Our results are shown in Table 3. It is evident that the SSD detector decreases re-id performance for all models as the SSD detectors seem to perform poorly on the PRW dataset. Regardless, our model outperforms both the BOW+XQDA baseline and the deep IDE feature embedding reported in [3] when the identical DPM person detector was used, by 2.2% and 7.4% in mAP and Rank-1, respectively. Note, the improved deep IDE$_{det}$ embedding of [3] outperforms DLDP by 2.9% and 2.3% in mAP and Rank-1 accuracy. However, this is due to its pretraining for person classification on the test domain data resulting in less false positive detections. In all experiments DLDP outperforms the baseline model, demonstrating the effectiveness of domain perceptive model selection.

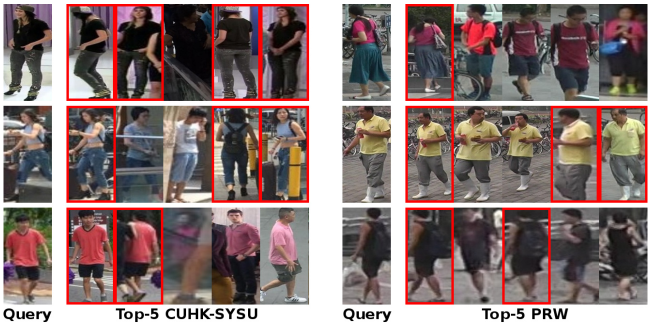

Fig. 3. Qualitative results of our approach. Correct results are framed in red. Note that most incorrect results are visually similar to the query.

5. CONCLUSION

In this work, we presented a novel approach to domain sensitive person re-identification by deep learning *without* the need for additional training data from the target (test) domains. Our evaluations on two latest benchmarks demonstrate clearly that the proposed DLDP model outperforms the state-of-the-art without use of test domain data and is even competitive to models trained with test domain data.

Acknowledgements: The research travel grant of the Karlsruhe House of Young Scientists (KHYS).

6. REFERENCES

[1] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, *Person re-identification*, vol. 1, Springer, 2014.

[2] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, 2016.

[3] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian, "Person re-identification in the wild," *arXiv preprint arXiv: 1604.02531*, 2016.

[4] Dong Yi, Zhen Lei, Shengcai Liao, Stan Z Li, et al., "Deep metric learning for person re-identification.," in *ICPR*, 2014.

[5] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016.

[6] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng, "An enhanced deep feature representation for person re-identification," in *WACV*, 2016.

[7] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, 2015.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[10] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[11] Lin Wu, Chunhua Shen, and Anton van den Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[12] Svebor Karaman, Giuseppe Lisanti, Andrew D Bagdanov, and Alberto Del Bimbo, "Leveraging local neighborhood topology for large scale person re-identification," *Pattern Recognition*, vol. 47, 2014.

[13] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *ICCV*, 2015.

[14] Ryan Layne, Timothy M Hospedales, and Shaogang Gong, "Domain transfer for person re-identification," in *ARTEMIS*, 2013.

[15] Lianyang Ma, Xiaokang Yang, and Dacheng Tao, "Person re-identification over camera networks using multi-task distance metric learning," *TIP*, vol. 23, 2014.

[16] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang, "Cross-scenario transfer person re-identification," *TCSVT*, 2015.

[17] Andy J Ma, Jiawei Li, Pong C Yuen, and Ping Li, "Cross-domain person reidentification using domain adaptation ranking svms," *TIP*, vol. 24, 2015.

[18] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016.

[19] Yang Hu, Dong Yi, Shengcai Liao, Zhen Lei, and Stan Z Li, "Cross dataset person re-identification," in *ACCV*, 2014.

[20] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *AVSS*, 2015.

[21] Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino, "The hda+ data set for research on fully automated re-identification systems," in *ECCV*, 2014.

[22] Chen Change Loy, Tao Xiang, and Shaogang Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *IJCV*, vol. 90, 2010.

[23] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *HGBU*. ACM, 2011.

[24] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011.

[25] Wei-Shi Zheng, Shaogang Gong, and Xiang Tao, "Associating groups of people," in *BMVC*, 2009.

[26] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011.

[27] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.

[28] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara, "Sarc3d: a new 3d body model for people tracking and re-identification," in *ICIAP*, 2011.

[29] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," *CVPR*, 2016.

[31] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," in *ICML*, 2016.

[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed, "Ssd: Single shot multibox detector," *ECCV*, 2016.

[33] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*. 2012.

[35] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, 2010.

[36] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.