One-Shot Metric Learning for Person Re-identification

Sławomir Bąk Peter Carr
Disney Research
Pittsburgh, PA, USA, 15213
{slawomir.bak,peter.carr}@disneyresearch.com

Abstract

Re-identification of people in surveillance footage must cope with drastic variations in color, background, viewing angle and a person's pose. Supervised techniques are often the most effective, but require extensive annotation which is infeasible for large camera networks. Unlike previous supervised learning approaches that require hundreds of annotated subjects, we learn a metric using a novel one-shot learning approach. We first learn a deep texture representation from intensity images with Convolutional Neural Networks (CNNs). When training a CNN using only intensity images, the learned embedding is color-invariant and shows high performance even on unseen datasets without fine-tuning. To account for differences in camera color distributions, we learn a color metric using a single pair of ColorChecker images. The proposed one-shot learning achieves performance that is competitive with supervised methods, but uses only a single example rather than the hundreds required for the fully supervised case. Compared with semi-supervised and unsupervised state-of-the-art methods, our approach yields significantly higher accuracy.

1. Introduction

Person re-identification is the task of finding the same individual across a network of cameras. A successful algorithm must cope with significant appearance changes caused by variations in color, background, camera viewpoint and a person's pose. Most successful state-of-the-art approaches employ supervised learning [14, 28, 32–34, 36, 62] and require hundreds of labeled image pairs of people across each camera pair. Novel deep architectures [2, 11, 55] can outperform these approaches, but training them from scratch requires thousands of labeled image pairs. Fine-tuning for target camera pairs [60] may help to decrease the amount of required training data to hundreds of image pairs. However, annotating hundreds of subjects in each camera pair is still tedious and does not scale to real-world networks. To overcome this issue, semi-supervised and unsupervised methods have been proposed [15, 26, 49, 52, 58]. Unfortunately, without labeled data, they usually look for feature invariance, which often reduces discriminativity and specificity (inability to adapt to camera-pair-specific changes). This makes them uncompetitive with supervised techniques. As a result, unsupervised and semi-supervised methods have received little attention in the research community because practicality and scalability have not been the main concern in current benchmark datasets (often limited to a small number of cameras).

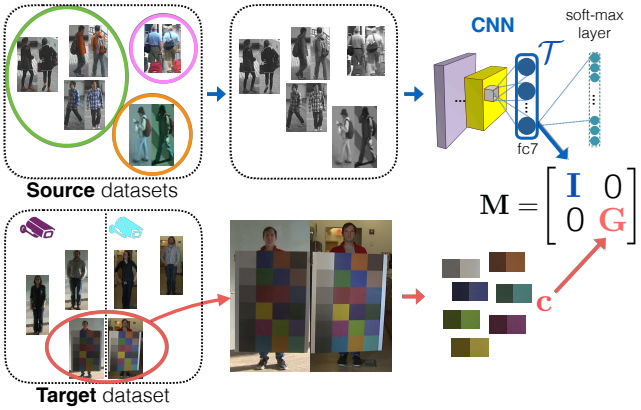


Figure 1: **One-Shot Metric Learning**. We split metric M into texture and color components. Deep texture features $\mathcal{T}$ are trained with CNN on *intensity* images to enforce color invariance without having to fine-tune. Joint learning on multiple **source** datasets (labeled data) increases good generalization under the Euclidean distance (identity matrix **I**). We adapt for color differences specific to **target** camera pair (unlabelled data) using a single image of color chart and learning color metric **G** for patch color features **c**.

In this paper, we propose a metric learning approach that scales to large camera networks by employing techniques similar to one-shot-learning [16]. We assume the metric learned for a pair of cameras can be split into texture and color components (see Fig. 1). For texture, we learn a color-invariant deep representation $\mathcal{T}$ that has good generalization abilities without fine-tuning. We can achieve this if we

use only *intensity* images and train a single CNN through a challenging multi-classification task on multiple datasets. In contrast, a CNN learned on color images would most likely require fine-tuning when testing [60], since the training dataset would have to be extremely large to cover all possible inter-camera color variations. Fine-tuning still requires a lot of training data, precluding its use with large camera networks. Instead, we incorporate color into our model using handcrafted color features that are independent of texture, and we learn a color metric **G** for each camera pair using a novel one-shot learning formulation. This strategy only requires a single example per camera, making it feasible for large networks. To account for specific differences in camera color distributions, we densely sample patches on registered images of a Macbeth ColorChecker chart [37] and learn a Mahalanobis metric that directly models the relationship between color features across a pair of cameras. Our contributions are:

- We split a metric for person re-identification into texture and color components. The metric is then learned on the target camera pair by a novel one-shot metric learning approach.
- Deep texture features are learned using only intensity images, thus ensuring invariance to color changes. Such features show high performance on unseen datasets without fine-tuning and are very competitive with semi- and unsupervised state-of-the-art methods.
- We adapt for color differences across cameras by learning a metric locally for patches using a single pair of images of a ColorChecker chart.
- Spatial variations in a person's appearance are incorporated into the color metric by explicitly modeling background distortions across cameras. When computing a distance between two images, we accommodate pose misalignments by defining a linear patch assignment problem, thus allowing patches to perturb their locations.

We conduct extensive experiments on five benchmark datasets. The results illustrate that by combining our deep texture features with a color metric trained using a single pair of images, we achieve very competitive performance with metrics learned from hundreds of examples. We outperform semi-supervised and unsupervised approaches and establish a new state of the art for scalable solutions for re-identification.

2. Related work

Supervised re-identification Most successful person re-identification techniques are based on *supervised* learning. They usually employ metric learning [2, 14, 28, 32, 33, 62] that uses training data to search for effective distance functions to compare people across different cameras. Many supervised machine learning algorithms have been considered for learning a robust metric. This includes feature selection by Adaboost [21], feature ranking by RankSVMs [45] and feature learning by convolution neural networks [2, 32, 51, 55, 60]. Although these deep convolution neural networks can be very effective, they often require thousands of image pairs to pre-train the architecture and hundreds of image pairs to fine-tune the network to a particular camera pair [55, 60]. To cope with insufficient data, data augmentation often has to be employed together with triplet embeddings [11].

Among all of these metric learning approaches, Mahalanobis distance functions [13, 22, 28, 54] received the most attention in the re-identification community [8]. Köstinger *et al*. [28] proposed a very effective and efficient KISS metric learning that uses a statistical inference based on a likelihood-ratio test of two Gaussian distributions modeling positive and negative pairwise differences between features. As this learning has an effective closed-form solution, many approaches have extended this work by introducing discriminative linear [34, 41] and non-linear [40, 56] subspace embeddings. Mahalanobis-like metric learning usually requires less training data than deep models (*i.e*. hundreds of labeled image pairs).

Recently, a trend of learning similarity measures for patches [4, 47, 48, 64] has emerged. Bak *et al*. [4] shows that learning metrics for patches might also effectively multiply the amount of training data (multiple patches may share the same metric). As a result, patch metrics can be learned on smaller amounts of labeled images (*e.g*. using 60 image pairs to infer an effective metric). However, annotating 60 subjects in each camera pair still does not scale to real-world scenarios, where a moderately-sized surveillance camera network can easily have hundreds of cameras.

Unsupervised re-identification Semi-supervised and unsupervised techniques have been proposed to avoid the scalability issue. Unsupervised approaches often focus on designing handcrafted features [5, 7, 12, 15, 53] that should be robust to changes in imaging conditions. One can further weight these features by incorporating unsupervised salience learning [52, 58] that looks for features that are far from the common distribution. Transfer learning has also been applied to re-identification [25, 63]. These methods learn the model using large labeled datasets (*e.g*. fashion photography datasets [49]) and transfer the discriminative knowledge to the unlabeled target camera pair.

Dictionary learning and sparse coding [1, 19, 35] have also been studied in context of re-identification. Dictionary learning derives from unsupervised settings, thus it can directly be applied to utilize unlabeled data to learn camera-invariant representations. To keep the dictionary discrim-

inative, graph Laplacian regularization is often introduced either to keep visually similar people close in the projected space [26, 27] or to perform cross-dataset transfer by multi-task learning [42]. Although the Laplacian regularization significantly helps, it is not sufficient to fully explore the discriminative space. As dictionary learning is prone to focus on invariant representations, there is still a considerable performance gap relative to supervised learning approaches.

One-shot learning One-shot learning aims at learning a task from one or very few training examples [16]. Usually, it involves a knowledge transfer either by model parameters [17] or by shared features [6]. In this work, we propose a one-shot metric learning approach where one part of the metric (texture) is transferred directly to the target dataset. The second part (color) is learned using patch-based metric learning. In contrast to the existing approaches that learn a metric from images of people, we learn a color metric using a single pair of Macbeth ColorChecker chart images [37]. This effectively reduces the amount of training data to a single example.

3. Method

Mahalanobis metric learning generates a metric $\mathbf{M}$ that measures the squared distance between feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j). \quad (1)$$

Köstinger [28] showed an effective closed-form solution (the KISS metric) to learn $\mathbf{M}$. In this paper, we propose to split the metric $\mathbf{M}$ into independent texture and color components, which is equivalent to enforcing a block diagonal structure: $\mathbf{M} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{G} \end{bmatrix}$, where the identity matrix $\mathbf{I}$ corresponds to the Euclidean distance between deep texture features (Sec. 3.1) and $\mathbf{G}$ is a color metric (Sec. 3.2) that is inferred using a single pair of images. In this context, we rewrite Eq. (1) and define the distance between two bounding box images i and j as

$$d^2(i,j) = (1-\gamma)||\mathcal{T}_i - \mathcal{T}_j||_2 + \gamma \Phi^2(\mathbf{c}_i, \mathbf{c}_j; \mathbf{G}), \quad (2)$$

where $\mathcal{T}_i$ and $\mathcal{T}_j$ are our deep texture features extracted after converting i and j to intensity images, $\mathbf{c}_i$ and $\mathbf{c}_j$ are color features extracted from color images and Φ is a Mahalanobis-like metric. Hyper-parameter $\gamma \in [0, 1]$ controls the importance of color relative to texture.

3.1. Texture dissimilarity

Perception of color is very susceptible to illumination changes. Even deep features [55, 60] learned on thousands of identities from multiple re-identification datasets require fine-tuning on unseen datasets. In contrast to these approaches, we are interested in a representation that does not require fine-tuning and can be applied directly to any camera pair. To achieve color-invariance, we drop the color information and convert all training images to single intensity channel images. We adopt the CNN model from [55] and train it from scratch using only intensity images to obtain highly robust color-invariant features for person re-identification. This model learns a set of high-level feature representations through challenging multi-class identification tasks, *i.e.*, classifying a training image into one of m identities. As the generalization capabilities of the learned features increase with the number of classes predicted during training [50], we need m to be relatively large (*e.g.* several thousand). As a result, we merge publicly available datasets into a single set of identities and train the network as joint single-task learning (JSTL) [55]. When it is trained to classify a large number of identities and configured to keep the dimension of the last hidden layer relatively low (*e.g.*, setting the number of dimensions for fc7 to 256 [55]), such CNNs form compact and highly robust texture representations for re-identification. In the rest of the paper, we refer to our neural network trained using only intensity images as JSTL^I and the features extracted from fc7 layer as $\mathcal{T}$. We found that directly using the Euclidean distance on such trained features is very effective; thus, we compute the dissimilarity score between $\mathcal{T}_i$ and $\mathcal{T}_j$ using ℓ_2 distance. In Sec. 4.2, we show that this texture representation has good generalization capabilities without fine-tuning and achieves competitive performance with semi- and unsupervised methods that utilize color information.

3.2. Color dissimilarity

In this section, we show how to learn a color metric using a single pair of images. We then allow this metric to vary spatially to cope with pose changes between images.

3.2.1 One-shot learning

Let $\mathbf{c}_i^A$ and $\mathbf{c}_j^B$ be the pair of color features extracted from two different cameras A and B. Typically [28], the space of pairwise differences $\mathbf{c}_{ij} = \mathbf{c}_i^A - \mathbf{c}_j^B$ is divided into positive pairwise set $\mathbf{c}_{ij}^+$ when i and j contain the same person, and $\mathbf{c}_{ij}^-$ otherwise. Learning the KISS metric involves computing two covariance matrices: Σ^+ for **positive pairwise differences** ($\Sigma^+ = (\mathbf{c}_{ij}^+)(\mathbf{c}_{ij}^+)^T$) and Σ^- for **negative pairwise differences** ($\Sigma^- = (\mathbf{c}_{ij}^-)(\mathbf{c}_{ij}^-)^T$). From the log-likelihood ratio, the Mahalanobis metric becomes $\mathbf{G} = (\Sigma^+)^{-1} - (\Sigma^-)^{-1}$ and measures the squared distance between two features $\mathbf{c}_i$ and $\mathbf{c}_j$

$$\Phi^2(\mathbf{c}_i, \mathbf{c}_j; \mathbf{G}) = (\mathbf{c}_i - \mathbf{c}_j)^T \mathbf{G} (\mathbf{c}_i - \mathbf{c}_j) \quad (3)$$

$$= \mathbf{c}_{ij}^T \left[(\Sigma^+)^{-1} - (\Sigma^-)^{-1} \right] \mathbf{c}_{ij}. \quad (4)$$

Covariance Σ^-: In practice, a set of negative examples can be generated by randomly selecting subjects' features from cameras A and B [28]. Even in the rare circumstance where a randomly generated pair of features corresponds to

the same individual, the odds of this happening frequently are nearly impossible.

Covariance Σ^+: Supervision is required for obtaining the set of positive pairwise examples $\mathbf{c}_{ij}^+$, thus computing Σ^+. We design separate foreground and background terms to facilitate learning. Let color features extracted in camera A be

$$\mathbf{c}_i^A = \mu_i + \sigma_i^A + \epsilon_i^A, \tag{5}$$

where μ_i is an implicit variable that refers to the i-th identity, σ_i^A denotes variations of μ_i, and ϵ_i^A corresponds to background distortion. The corresponding feature extracted for the same individual from camera B is $\mathbf{c}_j^B = \mu_i + \sigma_j^B + \epsilon_j^B$ (where $\mu_j = \mu_i$ because it is the same identity). Most approaches ignore foreground/background separation and assume metric learning will learn to identify and discard background features. In contrast, we explicitly model background distortions by ϵ. Computing positive pairwise differences we obtain

$$\begin{aligned}\mathbf{c}_{ij}^+ &= \mathbf{c}_i^A - \mathbf{c}_j^B \\ &= \sigma_i^A - \sigma_j^B + \epsilon_i^A - \epsilon_j^B \\ &= \Delta\sigma_{ij} + \Delta\epsilon_{ij}.\end{aligned} \tag{6}$$

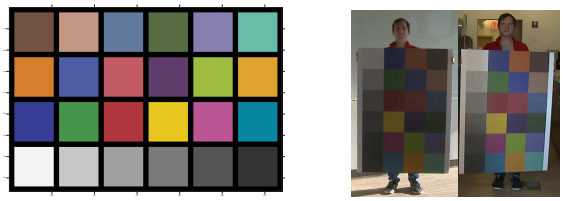
We assume that $\Delta\sigma$ and $\Delta\epsilon$ follow two different independent Gaussian distributions $\mathcal{N}(0, \Sigma_\sigma)$ and $\mathcal{N}(0, \Sigma_\epsilon)$, where Σ_σ and Σ_ϵ are unknown covariance matrices. The covariance of positive pairwise differences then becomes

$$\Sigma^+ = \Sigma_\sigma^+ + \Sigma_\epsilon^+. \tag{7}$$

To compute Σ_ϵ^+, we only need background images for a particular camera pair; thus, this information can be acquired without human supervision. For computing Σ_σ^+, we propose to use a ColorChecker calibration chart that holds information on color distribution in a given camera.

ColorChecker for Re-ID Driven by the idea that a good metric can be computed on the level of patches [4, 48], we design a new ColorChecker chart in such a way that corresponding patches across cameras can be used as different data points to compute $\mathbf{c}_{ij}^+$, thus obtaining Σ_σ^+ (Eq. 7). The standard Macbeth **ColorChecker** Chart [37] (see Fig. 2(a)) consists of color patches similar to natural objects, such as human skin, foliage, and flowers. The chart was designed to evaluate color reproduction processes by comparing the resulting images to the original chart.

Our design of ColorChecker chart for re-identification is based on insights from recent patch-based re-identification methods [4, 47, 48, 64]. The patch size matches the size of patches used for the re-id problem, and we removed the thin black borders to enable random sampling of the board (see Fig. 2(b)). This allows us to sample the space of color differences more effectively by exploring more points in the $\mathbf{c}_{ij}^+$ distribution (*e.g.*, combinations of different colors).



(a) (b)

Figure 2: Macbeth ColorCheckers (a) orignal [37]; (b) our ColorChecker for re-identification.

3.2.2 Spatial variations

Patch-based approaches [4, 48] generally perform better when metrics are allowed to vary spatially. Intuitively, regions with statistically different amounts of background distortion should have different metrics (*e.g.* patches in the leg region might contain more background pixels than patches at the torso). Let us assume that a bounding box image is divided into N patches. For a patch location n, we incorporate spatial variations into our model by redefining the Gaussian distribution of $\Delta\epsilon$ to be $\mathcal{N}(0, \alpha^{(n)}\Sigma_\epsilon)$, where $\alpha^{(n)}$ corresponds to the amount of environmental/background distortions and depends on the location n of the feature difference $\mathbf{c}_{ij}$ relative to the full bounding box of the detected person. As a result, Eq. 7 becomes

$$\Sigma^{+(n)} = \Sigma_\sigma^+ + \alpha^{(n)}\Sigma_\epsilon^+. \tag{8}$$

We usually expect $\alpha^{(n)}$ to be detector-dependent (based on how precisely the detector can generate a tight bounding box). We learn $\alpha^{(n)}$ using an auxiliary dataset. Let $\Sigma_R^{+(n)}$ be a covariance of positive pairwise differences computed from patches at location n using annotated individuals. We can learn $\alpha^{(n)}$'s by solving N objectives

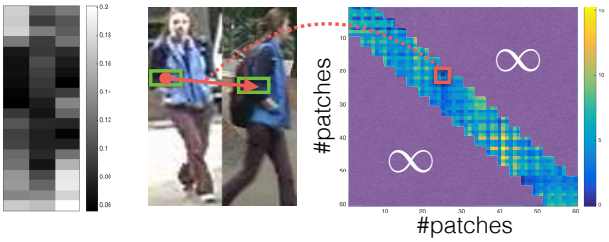
$$\alpha^{(n)} = \arg\min_\alpha ||\Sigma_\sigma^+ + \alpha\Sigma_\epsilon^+ - \Sigma_R^{+(n)}||_F : \alpha \in (0, 1), \tag{9}$$

for $n = 1 \ldots N$. We learn $\alpha^{(n)}$'s using annotated image pairs from the CUHK03 dataset and assume them to be fixed across all evaluation datasets (see Fig 3(a)). Note that larger amounts of background pixels yield higher values of α's (*e.g.* in head and leg regions). As a result, Φ and $\mathbf{G}$ from Eq. (3) become location dependent

$$\Phi^2(\mathbf{c}_i, \mathbf{c}_j; \mathbf{G}^{(n)}) = (\mathbf{c}_i - \mathbf{c}_j)^T\mathbf{G}^{(n)}(\mathbf{c}_i - \mathbf{c}_j), \tag{10}$$

$$\mathbf{G}^{(n)} = (\Sigma_\sigma^+ + \alpha^{(n)}\Sigma_\epsilon^+)^{-1} - (\Sigma^-)^{-1}. \tag{11}$$

Deformable model: In addition to spatially varying metrics, the correspondence between patches can also vary spatially. Because of pose changes, features extracted on a fixed grid may not correspond even though it is the same person. Therefore, patch-based methods [4, 48] often allow patches to adjust their locations when comparing two bounding box images. In [4], a deformable model consisted of spring constraints that controlled the relative place-



(a) (b)

Figure 3: Spatial variations: (a) learned background distortion coefficients $\alpha^{(n)}$; (b) $N \times N$ cost matrix, which is used as an input to the Hungarian algorithm for finding optimal patch correspondence.

ment of patches. These spring constraints were learned directly from data using structural SVMs. [47] assumed the correspondence structure to be fixed and learned it using a boosting-like approach. Instead, we define the patch correspondence task as a linear assignment problem. Given N patches from bounding box image i and N patches from bounding box image j we create a $N \times N$ cost matrix that contains patch similarity scores within a fixed neighborhood (see Fig 3(b)). To avoid patches freely changing their location, we introduce a global one-to-one matching constraint and solve a linear assignment problem

$$\mathbf{\Omega}^*_{ij} = \arg\min_{\mathbf{\Omega}_{ij}} \left(\sum_{n=1}^{N} \Phi^2(\mathbf{c}_i^{\mathbf{\Omega}_{ij}(n)}, \mathbf{c}_j^n; \mathbf{G}^{(n)}) + \Delta\big(\mathbf{\Omega}_{ij}(n), n\big) \right),$$
$$\text{s.t.} \quad \Delta\big(\mathbf{\Omega}_{ij}(n), n\big) = \begin{cases} \infty, & \eta(\mathbf{\Omega}_{ij}(n), n) > \delta; \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where $\mathbf{\Omega}_{ij}$ is a permutation vector mapping patches $\mathbf{c}_i^{\mathbf{\Omega}_{ij}(n)}$ to patches $\mathbf{c}_j^n$ and $\mathbf{\Omega}_{ij}(n)$ and n determine patch locations, $\Delta(\cdot,\cdot)$ is a spatial regularization term that constrains the search neighborhood, where η corresponds to distance between two patch locations and threshold δ determines the allowed displacement (different δ's are evaluated in Fig 7(a)). We find the optimal assignment $\mathbf{\Omega}^*_{ij}$ (patch correspondence) using the Kuhn-Munkres (Hungarian) algorithm [29]. This yields the color dissimilarity:

$$\sum_{n=1}^{N} \Phi^2(\mathbf{c}_i^{\mathbf{\Omega}^*_{ij}(n)}, \mathbf{c}_j^n; \mathbf{G}^{(n)}). \tag{13}$$

3.3. Total dissimilarity

By incorporating patches, Eq. (2) becomes

$$d^2(i,j) = (1-\gamma)||\mathcal{T}_i - \mathcal{T}_j||_2 + \gamma\left(\sum_{n=1}^{N} \Phi^2(\mathbf{c}_i^{\mathbf{\Omega}^*_{ij}(n)}, \mathbf{c}_j^n; \mathbf{G}^{(n)}) \right). \tag{14}$$

In the next section, we extensively evaluate both texture and color components as well as hyper-parameter γ.

Figure 4: Sample images from the **CCH** dataset: the top and bottom lines correspond to images from different cameras; columns illustrate the same person and the last column shows images of our ColorChecker chart.

4. Experiments

We carried out experiments on 5 datasets: **VIPeR** [20], **iLIDS** [61], **CUHK01** [31], **PRID2011** [23] and our new dataset, **CCH**. To learn a texture representation (fc7 of JSTL^I) and $\alpha^{(n)}$'s, we additionally used **CUHK03** [32]. Re-identification results are reported using the CMC curve [20] and its rank-1 accuracy. The CMC curve provides the probability of finding the correct match in the top r ranks.

4.1. Datasets and evaluation protocols

CCH (Color**CH**ecker) is our new dataset that consists of 23 individuals with 3379 images registered by two cameras in significantly different lighting conditions (see Fig. 4). A single pair of images of our ColorChecker chart was used to compute Σ^+_σ.

VIPeR [20] is one of the most popular person re-identification datasets. It contains 632 image pairs of pedestrians captured by two outdoor cameras. VIPeR images contain large variations in lighting conditions, background and viewpoint (see Fig. 5(a)).

CUHK01 [31] contains 971 people captured with two cameras. The first camera captures the side view of pedestrians and the second camera captures the front or back view (see Fig. 5(b)).

i-LIDS [61] consists of 476 images with 119 individuals. The images come from airport surveillance cameras. This dataset is very challenging because there are many occlusions due to luggage and crowds (see Fig. 5(c)).

PRID2011 [23] consists of person images recorded from two different static surveillance cameras. Characteristic challenges of this dataset are significant differences in illumination (see Fig. 5(d)). Although there are two camera views containing 385 and 749 identities, respectively, only 200 people appear in both cameras.

CUHK03 [32] is one of the largest published person re-identification datasets. It contains 1467 identities, so it fits very well for learning the JSTL model [55]. We used this dataset as an auxiliary dataset for training both deep texture representation and background distortion coefficients.

Evaluation protocols We fixed the evaluation protocol across all datasets. For computing color dissimilarity, all

(a) VIPeR (b) CUHK01

(c) iLIDS (d) PRID2011

Figure 5: Re-identification datasets and their synthesized ColorCheckers. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person; the last column illustrates our manually generated ColorCheckers.

images of individuals are scaled to be 128×48 pixels and divided into a set of 12×24 overlapping patches with a stride of 6×12 pixels. This yields 60 patches per image. To extract color features $\mathbf{c}_i$, we concatenate Lab, HSV, YCbCr, LUV, and RGB histograms, each with 10 bins per channel, into the 150-dimensional color feature vector, and we reduce the dimensionality to 30 components using PCA. For texture, we convert images to a single intensity channel. To fit the JSTL architecture [55], we scale them to be 160×64 pixels. For evaluation, we generated probe/gallery images accordingly to the settings in [40,55]: VIPeR: 316/316; CUHK01: 486/486; i-LIDS: 60/60; PRID: 100/649 and CCH: 23/23. In all experiments, we follow a single shot setting [40]. To obtain background patches for learning Σ_ϵ, we run background segmentation [43] and keep the patches that do not intersect with the foreground mask. For iLIDS and CCH we extract background patches from frames without subjects. To capture camera illumination conditions we use the Colorchecker chart. In practice, it is better (and easier) to use a picture of an actual chart. However for comparison purposes with existing datasets, we synthesize the ColorCheckers (see Fig. 5). We first randomly select 24 image pairs and extract 2 patches from the upper and the lower body parts. We then select 35 patches for the ColorChecker, while trying to match colors from Macbeth Chart [37]. Labeling 35 patches compares favorably to previous supervised learning methods that needed hand labeling of hundreds of subjects across each camera pair. This procedure was repeated 10 times to minimize subjective bias. $\mathbf{c}_{ij}^{+}$ is generated by randomly sampling 500 locations of the ColorCheckers.

4.2. Texture invariance

In this experiment we used 5 datasets: CUHK03, CUHK01, VIPeR, iLIDS and PRID. Similar to [55], we propose a joint learning scheme for producing an effective generic feature representation. We divide each dataset into training, testing and validation sets. As JSTL requires a high number of identities, all training, testing and validation sets are then merged into single training, testing and validation sets for training a single CNN. Individually training each dataset is usually not effective due to insufficient data [55]. In Tab. 1, we report comparison of JSTL trained only on intensity images (JSTL^I*) with JSTL trained on color images (JSTL*), and we refer to this scenario as supervised learning (because the training split from the test dataset was included in the merged training set). * is used to highlight supervised methods. Compared to KISSME [28] for both color and intensity images, it is apparent that a single CNN is flexible enough to handle multiple dataset variations. Learning on color images, we achieved better performance in this supervised setting.

	Method	VIPeR	CUHK	iLIDS	PRID	rel. perform. drop min	rel. perform. drop max	rel. perform. drop avg
intensity	JSTL^I*	15.8	50.6	44.1	35.0	-	-	
	JSTL$^I_{\text{LOO}}$	9.8	26.8	44.0	21.0	**0.2**	**47.0**	**31.3**
	Handcrafted	3.2	4.1	28.9	5.9	34.4	91.8	72.3
color	JSTL [55]*	35.4	62.1	56.9	59.0	-	-	
	JSTL$_{\text{LOO}}$	20.9	37.1	43.5	2.0	23.5	96.6	50.3
	KISSME [28]*	19.6	16.4	28.4	15.0	44.6	74.5	60.7
	Our	34.3	45.6	51.2	41.4	**3.1**	**29.8**	**17.3**

Table 1: CMC rank-1 accuracies, where * corresponds to the supervised methods. When training in leave-one-out (LOO) scenarios (unsupervised case), models trained only on intensity images have better generalization performance than models trained on color images (compare relative performance drop statistics). Our method is complementary to JSTL$^I_{\text{LOO}}$ and achieves significantly better accuracy than unsupervised methods and KISSME*, and it is comparable to supervised JSTL*.

However, as we are interested in generalization properties of this CNN (for unsupervised case), we also evaluate JSTL performance on unseen camera pairs. Similarly to leave-one-out (LOO) cross validation, we train CNNs from

scratch while entirely skipping images from the test camera pair (*e.g.* results of $\text{JSTL}^{I}_{\text{LOO}}$ in VIPeR column refers to JSTL trained using all datasets but VIPeR.). CUHK03 images were always included in the training phase. The right side of the table reports the performance drop statistics relative to the supervised JSTL($\frac{r_1^* - r_1}{r_1^*}$) for both intensity- and color-based models: *min*, *max* and *average* performance drop statistics across all datasets are provided. This experiment reveals that JSTL models trained on color images have significant relative performance drop, even up to 96.6% for the PRID dataset (*i.e.* rank-1 accuracy decreased from 59% to 2%). The average performance drop for color images is more than 50%. In contrast, for JSTL models trained using only intensity images, the performance drop is significantly lower and is even unnoticeable for some datasets (*e.g.*, iLIDS rank-1 dropped from 44.1% to 44.0%). This implies that models trained only on intensity images are more invariant to camera changes. $\text{JSTL}^{I}_{\text{LOO}}$ achieves reasonable performance without fine-tuning and is very competitive with the supervised KISSME [28] that uses color information, outperforming it on 3 of 4 datasets.

Intuitively, if we would have a large amount of data covering all possible color transfer functions, we should be able to learn features that have good generalization capabilities. In practice, with limited training data, our results indicate that it is more effective to learn deep texture representation using only intensity images and adapt to camera-pair specific color changes using the proposed one-shot learning (the last row in Tab. 1). Our approach significantly outperforms JSTL_{LOO} and KISSME and achieves comparable performance to its supervised counterpart – JSTL*.

Furthermore, to compare it with standard handcrafted texture descriptors, we concatenate HOG, LBP and SIFT features [58] extracted on a dense patch layout and compute image similarities using ℓ_2. From the results, it is apparent that $\text{JSTL}^{I}_{\text{LOO}}$ outperforms handcrafted features by a large margin on all datasets, which demonstrates the effectiveness of learning a set of generic deep texture features. As a result, we use $\text{JSTL}^{I}_{\text{LOO}}$ as our $\mathcal{T}_i$ descriptor.

4.3. Color calibration

Inter-camera color variation is an important problem for multi-camera systems. Standard approaches either (1) pursue *color constancy* (*i.e.*, perceiving the same color under different illuminations) and perform normalization techniques [18, 24, 30, 46] or (2) search for pair-wise mappings that are inferred from image pairs, *e.g.*, a pair of Macbeth ColorCheckers [3]. We compare our color metric learning to both groups of methods on the CCH dataset (now without the deep texture component, *i.e.* $\gamma = 1$ in Eq. (14)). The first group includes: histogram equalization (HQ) [24], multi-scale retinex with color restoration (MSRCR) [46], grey world normalization (GREY) [18] and log-chromaticity (LOG) [30]. The second group, which employs ColorChecker images, consists of: scene-specific color calibration (SSCC) [3], inter-camera color calibration (ICC) [44], and 3D Thin-plate smoothing spline (TPS) [9, 38]. A comparison in Fig. 6 between the two groups reveals that the performance of the second group (indicated by solid lines) is generally higher. It is also apparent that our color metric learning significantly outperforms all color calibration methods. Compensating for background distortions helps (*e.g.*, in Eq. 11, we can set learned coefficients $\alpha = \alpha^{(n)}$, or neglect background modeling $\alpha = 0$, or assume $\max$ background covariance computed from the training data across all patches $\alpha = 1$) but the main improvement comes from learning the metric $\mathbf{G}$ for color features using statistical inference [28]. Our approach yields significantly higher performance than standard approaches, which usually model color transfer either by 1D color histogram mappings [44] or low-rank matrix transforms [3].

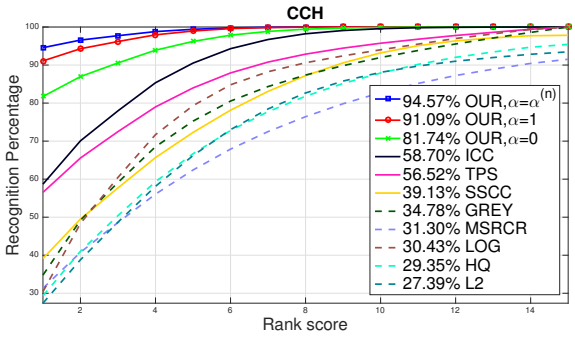


Figure 6: Performance comparison with standard color calibration techniques. Our approach outperforms other techniques by a large margin for all values of background distortion coefficients. The main improvement comes from learning metric $\mathbf{G}$.

4.4. Comparison to re-identification methods

Table. 2 reports the performance comparison of our one-shot metric learning with state-of-the-art approaches across 4 datasets. We report the results of unsupervised, semi-supervised and supervised approaches. Semi-supervised approaches usually assume the availability of one third of the training set. The #IDs column provides the average number of labeled identities used for training corresponding models. Our method outperforms all semi- and unsupervised methods on all datasets, and it achieves maximum improvement on the PRID dataset. We improve the state of the art by more than 16% on rank-1 accuracy compared to the previous best reported result, including results from unsupervised GL [26] and semi-supervised TL-semi [42] approaches. Further, our approach achieves competitive performance with the best supervised methods that require hundreds of training examples. For example, our results on the PRID dataset outperform all supervised ap-

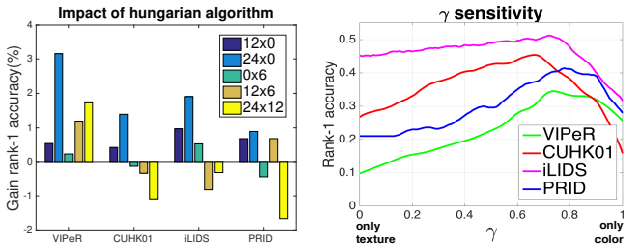
	METHOD	#IDs	VIPeR	CUHK01	iLIDS	PRID
semi/unsupervised	**Our**, $\alpha = \alpha^{(n)}$	1	**34.3**	**45.6**	**51.2**	**41.4**
	Our, $\alpha = 0$	1	30.1	39.6	49.9	31.9
	$\text{JSTL}^{I}_{\text{LOO}}$	0	9.8	26.8	44.0	21.0
	JSTL_{LOO}	0	20.9	37.1	43.5	2.0
	Null Space-semi [57]	80	31.6	-	-	24.7
	GL [26]	0	33.5	41.0	-	25.0
	DLLAP-un [27]	0	29.6	28.4	-	21.4
	DLLAP-semi [27]	80	32.5	-	-	22.1
	eSDC [58]	0	26.7	15.1	36.8	-
	GTS [52]	0	25.2	-	42.3	-
	SDALF [15]	0	19.9	9.9	41.7	16.3
	TSR [49]	0	27.7	23.3	-	-
	TL-un [42]	0	31.5	27.1	49.3	24.2
	TL-semi [42]	80	34.1	32.1	50.3	25.3
supervised	FT-JSTL+DGD [55]	2629	38.6	**66.6**	**64.6**	**64.0**
	KISSME [28]	240	19.6	16.4	28.4	15.0
	LOMO+XQDA [34]	240	40.0	63.2	-	26.7
	Mirror [10]	240	42.9	40.4	-	-
	Ensembles [40]	240	45.9	53.4	50.3	17.9
	MidLevel [59]	240	29.1	34.3	-	-
	DPML [4]	240	41.4	35.8	57.6	-
	kLDFA [56]	240	32.8	-	40.3	22.4
	DeepNN [2]	240	34.8	47.5	-	-
	Null Space [57]	240	42.2	64.9	-	29.8
	Triplet Loss [11]	240	47.8	53.7	60.4	22.0
	Gaussian+XQDA [36]	240	**49.7**	57.8	-	-

Table 2: CMC rank-1 accuracies. The best scores for un- and semi-supervised methods are shown in blue. Our approach performs the best among all these methods across all datasets. The best scores of supervised methods are highlighted in red. Our results are comparable with supervised methods that require hundreds or thousands of identified image pairs for training.

proaches except FT-JSTL+DGD [55]. This model was pre-trained on 2629 subjects, and hundreds of image pairs were used to fine-tune the model on the target dataset. In a real-world scenario, collecting these hundreds of images pairs might be very difficult, if not impossible. Our model needs only a single pair of images, which is a reasonable requirement for real-world deployments.

4.5. Model parameters

Deformable model Fig. 7(a) illustrates the impact of a deformable model on recognition accuracy. We also compare the effectiveness of different neighborhoods on the overall accuracy. In Eq. (12), we constrain the displacement of patches to $\delta_{\text{horizontal}} \times \delta_{\text{vertical}}$ number of pixels. Interestingly, allowing patches to move vertically ($\delta_{\text{vertical}} > 0$) generally decreases performance. We believe that this is due to the fact that images in all of these datasets were annotated manually and vertical alignment (from the head to the feet) of people in these images is usually correct. Allowing patches to move horizontally consistently improves the performance for all datasets. The highest gain in accuracy is obtained on the VIPeR dataset (+3.2%), which was originally designed for evaluating viewpoint invariance. This indicates that our linear assignment approach provides a reliable solution for pose changes.



(a) (b)

Figure 7: Models parameters: (a) comparison of different allowable neighborhoods (horizontal×vertical) when applying Hungarian algorithm for matching patches; (b) sensitivity of hyper-parameter γ.

The color importance Intuitively, it seems that color should hold the most discriminative information of a person's identity. Conversely, we find that by employing only the intensity channel, we can achieve a fairly strong baseline for person re-identification. Color, although discriminative, is very susceptible to illumination changes. Interestingly, it is not clear which has more impact on the final performance – our one-shot color metric learning or the deep texture representation. Compare two extremes in Fig. 7(b): using only texture $\gamma = 0$, and using only color $\gamma = 1$. Texture alone performs better than color alone on two datasets (iLIDS, CUHK01) but it is outperformed on two others (VIPeR, PRID). Combining texture and color components consistently increases the recognition accuracy in all datasets.

Computational complexity Eq. (12) requires solving Hungarian algorithm for relatively sparse 60×60 matrix (see Fig. 3(b)). Given k non-infinite entries in this matrix, we employed QuickMatch algorithm [39] that runs in linear time $\mathcal{O}(k)$. The deep texture feature extraction is the slowest part and it depends on the GPU architecture (*e.g.* on Tesla K80 VIPeR experiment takes 45s, with 39s spent on deep feature extraction).

5. Summary

Supervised re-identification approaches require hundreds of labeled image pairs to train effective models for each camera pair. This does not scale to real-world scenarios where the number of cameras in a surveillance network could be large. In this paper, we presented a novel one-shot learning approach that achieves competitive performance with the best supervised learning approaches, but only requires a single image from each camera for training. We assume a metric can be split into independent color and texture components without loss of performance. For texture, we learn deep color-invariant features that can be directly applied to unseen camera pairs without fine-tuning. Color variations for specific camera pairs are captured by sampling patches on registered images of a ColorChecker chart and learning a color metric for patches. Our method leads to new state-of-the-art performance in practical and scalable solutions for re-identification.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. k -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, Nov 2006. 2

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2, 8

[3] D. Akkaynak, T. Treibitz, B. Xiao, U. A. Gürkan, J. J. Allen, U. Demirci, and R. T. Hanlon. Use of commercial off-the-shelf digital cameras for scientific data acquisition and scene-specific color calibration. *J. Opt. Soc. Am. A*, 31(2):312–321, Feb 2014. 7

[4] S. Bak and P. Carr. Person re-identification using deformable patch metric learning. In *WACV*, 2016. 2, 4, 8

[5] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010. 2

[6] E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *CVPR*, 2005. 3

[7] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *ICPR*, pages 1413–1416, 2010. 2

[8] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709, 2013. 2

[9] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, June 1989. 7

[10] Y.-C. Chen, W.-S. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *IJCAI*, 2015. 8

[11] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, June 2016. 1, 2, 8

[12] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 2

[13] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 2

[14] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, pages 501–512, 2010. 1, 2

[15] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 2, 8

[16] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. 1, 3

[17] M. Fink. Object classification from a single example utilizing class relevance metrics. In *NIPS*, 2004. 3

[18] G. D. Finlayson, B. Schiele, and J. L. Crowley. Comprehensive colour image normalization. In *ECCV*, 1998. 7

[19] S. Gao, I. W. H. Tsang, L. T. Chia, and P. Zhao. Local features are not lonely - laplacian sparse coding for image classification. In *CVPR*, 2010. 2

[20] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *PETS*, 2007. 5

[21] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2

[22] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 2

[23] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 5

[24] S. D. Hordley, G. D. Finlayson, G. Schaefer, and G. Y. Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38, 2005. 7

[25] J. Hu, J. Lu, and Y.-P. Tan. Deep transfer metric learning. In *CVPR*, 2015. 2

[26] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised l1 graph learning. In *ECCV*, 2016. 1, 3, 7, 8

[27] E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015. 3, 8

[28] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 2, 3, 4, 6, 7, 8

[29] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 1955. 5

[30] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *TPAMI*, 2013. 7

[31] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 5

[32] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 5

[33] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 1, 2

[34] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 2, 8

[35] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, June 2014. 2

[36] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 8

[37] C. S. McCamy, H. Marcw, and J. G. Davidson. A color rendition chart. *Journal of Applied Photographic Engineering*, 1976. 2, 3, 4, 6

[38] P. Menesatti, C. Angelini, F. Pallottino, F. Antonucci, J. Aguzzi, and C. Costa. Rgb color calibration for quantitative image analysis: The 3d thin-plate spline warping approach. *Sensors*, 12(6):7063, 2012. 7

[39] J. B. Orlin and Y. Lee. QuickMatch: A very fast algorithm for the Assignment Problem. Technical Report WP 3547-93, Massachusetts Institute of Technology, 1993. 8

[40] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 2, 6, 8

[41] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 2

[42] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, June 2016. 3, 7, 8

[43] A. Perina, N. Jojic, M. Cristani, and V. Murino. Stel component analysis: Joint segmentation, modeling and recognition of objects classes. *IJCV*, 100(3):241–260, 2012. 6

[44] F. Porikli. Inter-camera color calibration by correlation model function. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–133–6 vol.3, Sept. 2003. 7

[45] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010. 2

[46] Z.-u. Rahman, D. J. Jobson, and G. A. Woodell. Retinex processing for automatic image enhancement. In *Proc. SPIE*, volume 4662, pages 390–401, 2002. 7

[47] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015. 2, 4, 5

[48] H. Sheng, Y. Huang, Y. Zheng, J. Chen, and Z. Xiong. Person re-identification via learning visual similarity on corresponding patch pairs. In *KSEM*, 2015. 2, 4

[49] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015. 1, 2, 8

[50] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, June 2014. 3

[51] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[52] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 1, 2, 8

[53] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, pages 1–8, 2007. 2

[54] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006. 2

[55] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1, 2, 3, 5, 6, 8

[56] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 2, 8

[57] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 8

[58] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2, 7, 8

[59] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 8

[60] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 1, 2, 3

[61] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 5

[62] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 1, 2

[63] W. S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, March 2016. 2

[64] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015. 2, 4