Person Depth ReID: Robust Person Re-identification with Commodity Depth Sensors

Nikolaos Karianakis
University of California, Los Angeles
Los Angeles, CA 90095
nikos.karianakis@gmail.com

Zicheng Liu
Microsoft Research
Redmond, WA 98052
zliu@microsoft.com

Yinpeng Chen
Microsoft Research
Redmond, WA 98052
yiche@microsoft.com

Stefano Soatto
University of California, Los Angeles
Los Angeles, CA 90095
soatto@cs.ucla.edu

ABSTRACT

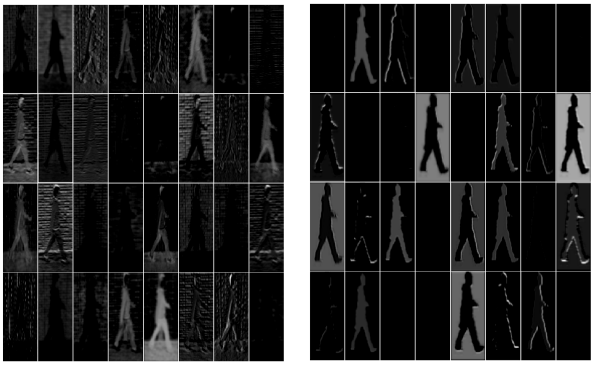
This work targets person re-identification (ReID) from depth sensors such as Kinect. Since depth is invariant to illumination and less sensitive than color to day-by-day appearance changes, a natural question is whether depth is an effective modality for Person ReID, especially in scenarios where individuals wear different colored clothes or over a period of several months. We explore the use of recurrent Deep Neural Networks for learning high-level shape information from low-resolution depth images. In order to tackle the small sample size problem, we introduce regularization and a hard temporal attention unit. The whole model can be trained end to end with a hybrid supervised loss. We carry out a thorough experimental evaluation of the proposed method on three person re-identification datasets, which include side views, views from the top and sequences with varying degree of partial occlusion, pose and viewpoint variations. To that end, we introduce a new dataset with RGB-D and skeleton data. In a scenario where subjects are recorded after three months with new clothes, we demonstrate large performance gains attained using Depth ReID compared to a state-of-the-art Color ReID. Finally, we show further improvements using the temporal attention unit in multi-shot setting.

KEYWORDS

Person re-identification, depth sensors, deep learning, temporal attention.

1 INTRODUCTION

Person re-identification is a fundamental problem in automated video surveillance and has attracted significant attention in recent years [5, 18, 59]. When a person is captured by cameras with non-overlapping views, or by the same camera but over many days, the objective is to recognize them across views among a large number of imposters. This is a difficult problem because of the visual ambiguity in a person's appearance due to large variations in illumination, human pose, camera settings and viewpoint. Additionally, re-identification systems have to be robust to partial occlusions and cluttered background. Multi-person association has wide applicability and utility in areas such as robotics, multimedia, forensics, autonomous driving and cashier-free shopping.



Person Color ReID [63] Person Depth ReID

Figure 1: Convolutional filter responses from "conv3" layer using the same frame from the TUM GAID data as input for both Person Color ReID [63] and the feature encoder f_{CNN} of Person Depth ReID, which is drawn in Fig. 3.

1.1 Related work

Existing methods of person re-identification typically focus on designing invariant and discriminant features [8, 16, 19, 30, 33, 34, 37, 66, 76], which can enable identification despite nuisance factors such as scale, location, partial occlusion and changing lighting conditions. In an effort to improve their robustness, the current trend is to deploy higher-dimensional descriptors [33, 35] and deep convolutional architectures [1, 31, 55, 60, 63, 68].

In spite of the ongoing quest for effective representations, it is difficult to deal with very large variations such as ultra wide-baseline matching and dramatic changes in illumination and image resolution, especially when having limited training data. As such, there is vast literature in learning discriminative distance metrics [13, 28, 32, 33, 38, 40, 42, 48, 58, 79] and discriminant subspaces [33–35, 49, 50, 64, 72]. Other approaches alleviate the problem of pose variability by explicitly accounting for spatial constraints of the human body parts [10] or by predicting the pose within a multi-shot setting [12]. Similarly, adjacency constrained salient region matching [74, 75] can help tackle the misalignment caused by large viewpoint and pose variation. In order to reduce the

intra-class variance while preserving the intrinsic graphical structure, Shi et al. [52] mined positive and negative samples of different difficulty and built graphical relationships to approximate geodesic distance for training convolutional neural networks. Kodirov et al. [27] followed unsupervised methodology to formulate a graph regularized dictionary learning model and efficient optimization algorithm for cross-view matching.

However, a key challenge to tackle within both distance learning and deep learning pipelines is the *small sample size* problem [11, 72], which is attributed to the lack of large-scale person re-identification benchmarks. New datasets have been released recently, such as CUHK03 [31] and MARS [77], a video extension of the Market-1501 dataset [78]. Their training sets are in the order of 20,000 positive samples, i.e. two orders of magnitude smaller than Imagenet [51] which has been successfully used for object recognition [29, 54, 57].

The small sample size problem is especially acute on the person re-identification algorithms which leverage temporal sequences [7, 20, 41, 65], as the feature dimensionality increases linearly in the number of frames that are accumulated compared to the single-shot representations. On the other hand, explicitly modeling temporal dynamics and using multiple frames help algorithms to deal with noisy measurements, occlusions, adverse poses and lighting.

Adding regularization, such as *Dropout* [21], to the layers where most parameters are concentrated like the fully-connected ones, is one step towards reducing the parameter space and allow learning models to have higher generalization capability. Xiao et al. [63] achieved state-of-the-art accuracy on many person re-identification benchmarks by designing a deep convolutional network, similar in nature to *GoogleNet* [57], and training it on the union of several available datasets. Additionally, they further improve their performance on individual datasets by introducing "domain-guided dropout", where the dropout rate for each neuron is adaptively set as a function of its activation rate in the training set.

Haque et al. [20] introduced a carefully designed *glimpse* layer in order to compress their 4D spatiotemporal input representation of 500-frame video from $\approx 2.5 \times 10^9$ elements to a feature vector size in the order of 1×10^6 elements. They provide the compressed vector as input to a 4D convolutional encoder, while the decision for the next glimpse location is made using a sparsification technique with a reinforcement learning objective within a recurrent attention framework. However, designing such a downsampling mechanism with the objective of minimizing the large input size without losing much information can be challenging. Our algorithm has several key differences with this work: First, we do not use any glimpse layer and there is no locator module, as our input module detects the human silhouette region, which is used in its entirety. Second, instead of a $4D$ convolutional autoencoder, our encoder is 3-dimensional and its input is one frame. Third, we design a temporal attention unit to estimate the weight of each frame, which regularizes the recurrence and affects the multi-shot evaluation.

Some recent works in natural language processing [9, 36] explore temporal attention in order to keep track of long-range structural dependencies. Yao et al. [67] in video captioning use a soft attention gate inside their Long Short-term memory decoder, so that they estimate the relevance of current features in the input video given all the previously generated words. Our algorithm is different from these approaches as we use a *hard attention* unit, which is not differentiable but can be learned with reinforcement learning.

In the literature there are RGB-based approaches which extract the binary silhouettes and estimate geodesic distances between body parts [26, 39, 71]. Also, depth-based methods that use measurements from 3D human skeleton data have emerged in order to infer anthropometric and human gait criteria [2, 3, 15, 44, 47]. In an effort to leverage the full power of depth data, recent methods use 3D point clouds to estimate motion trajectories and the length of specific body parts [23, 73]. It is worthwhile to point out that skeleton information is not always available. For example, the skeleton tracking in Kinect SDK can be ineffective when a person is in side view or the legs are not visible.

1.2 Motivation

On top of the above-mentioned challenges, RGB-based methods are challenged in scenarios with significant lighting changes and when the individuals change clothes. These factors can have a major influence on the effectiveness of a system that, for instance, is meant to track people across different areas of a building over several days where different areas of a building may have very different lighting conditions, the cameras may have different color balance, and a person may wear clothes of different colors. This is our *key motivation* for investigating representations that are insensitive to color information such as silhouettes from depth.

1.3 Contributions

Our contributions can be summarized as follows:

i) We explore the use of depth sensors for person re-identification under adverse conditions, such as cases where the subjects appear with different clothes over time, while still being robust to viewpoint variation, human pose and partial occlusion. We construct representations from depth, so that we enable feature learning in end-to-end fashion with deep convolutional neural networks. The learned representations are different in nature from those learned with RGB models (see Fig. 1). Our experiments (e.g., see Fig. 5) suggest that depth is an effective modality for this task.

ii) We tackle the small sample size problem in various ways: first, we customize the optimization for the depth modality and deploy dropout in the convolutional encoder and the recurrent element. Second, we use the time as regularizer, as the agent is a recurrent neural network. Third, we design a temporal hard attention unit, whose weights are learned with a reinforcement learning objective, and enables scalability over longer sequences. Fourth, initializing the encoder with a pre-trained RGB ReID model [63] provides multimodal data augmentation.

iii) We conduct an empirical study using three re-identification datasets. TUM-GAID database [22] is the largest one, including 305 persons. In the scenario where 32 subjects appear with different clothes after three months, our model achieves 6.2% higher top-1 and 23.6% higher top-5 accuracy compared to the top-performing RGB-based ReID method [63]. Next, we show further performance improvements when deploying recurrence and temporal attention, along with the head color information. We use the DPI-T dataset [20] with views from top to compare with Haque et al., who released this dataset and used an attention model with spatial

glimpse layer. Finally, in order to evaluate the effectiveness of our algorithm with more challenging partial occlusions, viewpoints, and human poses, we introduce the FaceBody dataset. It involves 57 subjects that walk and operate in a realistic meeting room scenario.

2 OUR METHOD

2.1 Input Representation

The input for our system is raw depth measurements from the Kinect V2 Sensor. The input data are depth images $\mathbf{D} \in \mathbb{Z}^{512\times424}$, where each pixel $D[i,j], i \in [1,\ldots,512], j \in [1,\ldots,424]$, contains the Cartesian distance, in millimeters, from the image plane to the nearest object at the particular (i,j) coordinate. In "default range" setting, $[0, 0.4m)$ and $(8.0m, \infty)$ ranges are classified as unknown measurements, $[0.4, 0.8)[m]$ as "too near", $(4.0, 8.0][m]$ as "too far" and $[0.8, 4.0][m]$ as "normal" values. We have a dedicated algorithm to crop a rectangle that surrounds the person. When skeleton tracking is effective, the *body index* $\mathbf{B} \in \mathbb{Z}^{512\times424}$ is provided by the Kinect SDK, where 0 corresponds to background and a positive integer i for each pixel belonging to the person i. Therefore, when the Body Index is available, there is no need to use tracking in order to effectively crop the person (see Sec. 3.6).

After extracting the person region $\mathbf{D_p} \subset \mathbf{D}$, the measurements within the "normal" region are normalized in the range $[1, 256]$, while the values from "too far" and "unknown" range are set as 256, and values within the "too near" range as 1. In practice, in order to avoid a concentration of the values near 256, whereas other values, say on the floor in front of the subject, span the remaining range, we introduce an offset $t_o = 56$ and normalize in $[1, 256 - t_o]$. This results in the "grayscale" representation $\mathbf{D_p^g}$. When the skeleton information is available, the body index $\mathbf{B_p} \subset \mathbf{B}$ is used as binary mask for background subtraction on the person depth region $\mathbf{D_p}$ before range normalization (see Fig. 2). Assuming that we crop person i, each pixel of $\mathbf{B_p}$ with body index value different from i is set to 256.

We also consider the binary representation $\mathbf{D_p^b}$, as "black-and-white silhouette", by thresholding on $t_b = 128$:

$$D_p^b(i,j) = \begin{cases} 1, & \text{if } D_p^g(i,j) < t_b \\ 128, & \text{if } D_p^g(i,j) >= t_b \end{cases} \quad (1)$$

for $(i,j) \in [1,512] \times [1,424]$. The average image is computed over the training set and is subtracted from each testing image.

2.2 Model

The problem is formulated as *sequential decision process* of an agent that performs human recognition from a partially observed environment via video sequences. At each time step, the agent observes the environment via depth camera, calculates a feature vector based on a deep Convolutional Neural Network (CNN) and actively infers the importance of the current frame for the re-identification task via a temporal attention unit. The weight that is estimated by the attention unit determines whether the hidden representation is updated or not, which subsequently affects the classification. This hidden representation is computed by a recurrent module, which is meant to model the temporal dynamics. Finally, the agent receives a reward based on the success or failure of its action at each step. The agent's objective is to maximize the sum of rewards over time.

Figure 2: The cropped color image (left), the grayscale depth representation $\mathbf{D_p^g}$ (center) and the result after background subtraction (right) using the body index information $\mathbf{B_p}$ from skeleton tracking.

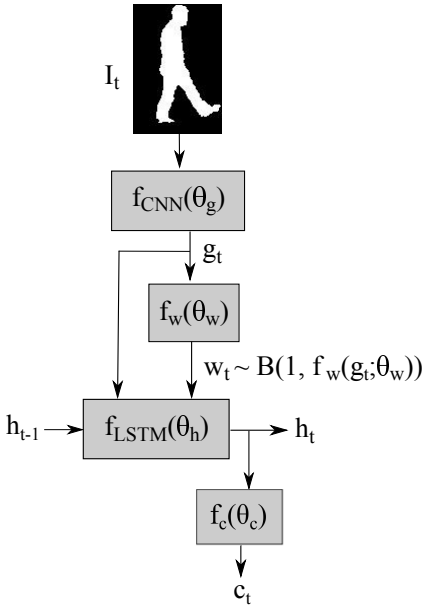


Figure 3: Model architecture: a recurrent deep neural network with temporal attention.

The agent, as well as its comprising modules, are described in the following paragraphs. An outline of the model is shown in Fig. 3.

2.2.1 Agent. Formally, the problem setup is a Partially Observable Decision Process (POMDP). The true state of the environment is unknown. The agent learns a stochastic policy $\pi((w_t, c_t)|s_{1:t}; \theta)$ with parameters $\theta = \{\theta_g, \theta_w, \theta_h, \theta_c\}$ that, at each step t, maps the history of past information $s_{1:t} = I_1, w_1, c_1, \ldots, I_{t-1}, w_{t-1}, c_{t-1}, I_t$ to a distribution of actions. Both actions contribute to the recognition task via the estimated frame weight w_t and class posterior c_t. The weight w_t is computed by the temporal attention unit, which

takes the current frame encoding g_t as input, while the classifier is attached on the RNN output h_t. The vector h_t maintains an internal state of the environment as a summary of past observations and is updated by the recurrent module $f_{LSTM}(\theta_h)$. Note that, for simplicity of notation, the input image at time t is denoted as I_t, but the exact input representation is the grayscale region D_p^g as described in Sec. 2.1. At each time step t, the agent receives a reward r_t, which equals to 1 when the frame is correctly classified and 0 otherwise.

2.2.2 Feature encoder $f_{CNN}(\theta_g)$. The first design choice pertains to choosing features that are robust to various image and human shape variations due to camera viewpoint, human pose, light conditions, noisy measurement and partial occlusion. Recent investigation for the best architecture for person re-identification [1, 31, 55, 60, 63, 68] has shown that the deep convolutional network introduced by Xiao et al. [63] has outperformed other approaches on several public datasets. This network uses batch normalization [24] and includes 3×3 convolutional layers [54], followed by 6 Inception modules [57], and 2 fully connected layers. We adopt this architecture because in addition to its effectiveness in RGB-based person re-identification, it allows us to initialize the parameters of Depth ReID with a pre-trained model, as a form of data augmentation.

We introduce two modifications in this network. We replace the top layer with a $256 \times N$ fully connected layer, where N is the number of subjects and depends on the dataset. The weights of this layer are initialized at random from a zero-mean Gaussian distribution with standard deviation 0.01. Additionally, we add dropout regularization between the fully-connected layers.

The model is trained to recognize the *identity* of a person by minimizing its cross-entropy loss, as is customary in other large-scale recognition tasks, such as face identification [56]. Afterwards, we remove the model's top layer and use the 256×1 vector as our feature encoding g_t.

2.2.3 Recurrent module $f_{LSTM}(\theta_h)$. We use Long Short-Term Memory (LSTM) element units as described in [70], which have been shown by Donahue et al. [14] to be effective in dealing with the vanishing and exploding gradients problem and modeling long-term dynamics for computer vision tasks. Assuming that $\sigma()$ is sigmoid, $g[t]$ is the input at time frame t, $h[t-1]$ is the previous output of the module and $c[t-1]$ is the previous cell, the implementation corresponds to the following updates:

$$i[t] = \sigma(W_{gi}g[t] + W_{hi}h[t-1] + b_i) \quad (2)$$

$$f[t] = \sigma(W_{gf}g[t] + W_{hf}h[t-1] + b_f) \quad (3)$$

$$z[t] = tanh(W_{gc}g[t] + W_{hc}h[t-1] + b_c) \quad (4)$$

$$c[t] = f[t] \odot c[t-1] + i[t] \odot z[t] \quad (5)$$

$$o[t] = \sigma(W_{go}g[t] + W_{ho}h[t-1] + b_o) \quad (6)$$

$$h[t] = o[t] \odot tanh(c[t]) \quad (7)$$

where W_{sq} is the weight matrix from source s to target q for each gate q, b_q are the biases leading into q, $i[t]$ is the input gate, $f[t]$ is the forget gate, $z[t]$ is the input to the cell, $c[t]$ is the cell, $o[t]$ is the output gate, and $h[t]$ is the output of this module. Finally, $x \odot y$ denotes the element-wise product of vectors x and y. Note that this LSTM does not use peephole connections between cell and gates.

2.2.4 Temporal attention unit $f_w(\theta_w)$. At each time step t the attention unit calculates the weight w_t of the image frame I_t, as the latter is represented by the feature encoding g_t. This module consists of a linear layer which maps the 256×1 vector g_t to one scalar, followed by Sigmoid non-linearity which squashes real-valued inputs to a $[0, 1]$ range. Next, the output of the module is defined by a Bernoulli random variable with probability mass function:

$$f(w_t; f_w(g_t; \theta_w)) = \begin{cases} f_w(g_t; \theta_w), & \text{if } w_t = 1 \\ 1 - f_w(g_t; \theta_w), & \text{if } w_t = 0 \end{cases} \quad (8)$$

During training, the weight w_t is chosen *stochastically* to be a binary value in $\{0, 1\}$. When $w_t = 1$, the current input g_t is forwarded through the LSTM. In case $w_t = 0$, the recurrent module is bypassed and the hidden representation from the previous frame is propagated to the current frame ($h_t := h_{t-1}$). During testing, the temporal unit acts deterministically and therefore $w_t = f_w(g_t; \theta_w)$.

This stochastic procedure introduces noise at the frame level during training, which is analogous to dropout regularization, but with a data-driven Bernoulli parameter instead. The probability of dropping a frame is controlled by the parameter $p = f_w(g_t; \theta_w)$, which ensures learning better models, as shown empirically in Sec. 3.7. Frames that the encoder is more confident to classify correctly are less likely to be dropped, as opposed to frames with a low-confidence encoder. This behavior is learned via reinforcement learning as explained in Sec. 2.3.2. An example sequence with the inferred Bernoulli parameter p for each frame is shown in Fig. 6.

2.2.5 Classifier $f_c(\theta_c)$. The classifier consists of a fully connected layer and Softmax, which map the 256×1 hidden vector h_t to the posterior class vector c_t with length N depending on the dataset. We use dropout regularization between the hidden vector and the classifier.

2.3 Training

In our experiments we pre-train the parameters of the feature encoder θ_g before attaching the RNN and the attention module and train the whole model in end-to-end-fashion. However, the entire architecture can be trained from scratch end to end. In the following subsections we describe the training process for the encoder and the recursive model with attention using a hybrid supervised loss.

2.3.1 Training the encoder $f_{CNN}(\theta_g)$. Deploying popular training techniques [6] with depth data needs careful consideration regarding the optimization process, as the dataset size is typically limited and the representations are of different nature than those that are color-based (see Fig. 2). We found empirically that stochastic gradient descent with modest base learning rate and low momentum can consistently converge to a good local minimum.

Optimization. Formally, stochastic gradient descent updates the model's weights w using a linear combination of the negative gradient of the loss $Q(z, w)$ for input z with respect to the weights w and the previous weight update v. The learning rate γ and the momentum μ are the coefficients of these two terms, respectively.

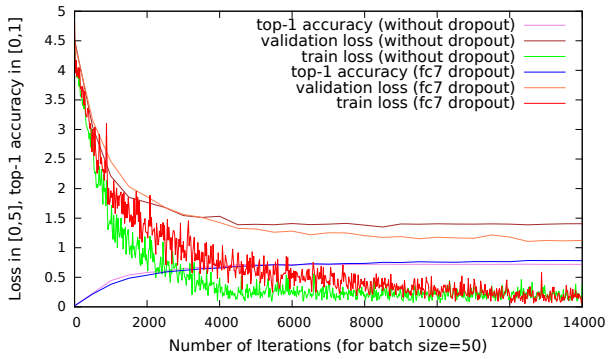


Figure 4: The encoder convergence on FaceBody data.

At time t the update is:

$$v_{t+1} = \mu v_t - \gamma_t \nabla_w Q(z_t, w_t) \tag{9}$$

$$w_{t+1} = w_t + v_{t+1} \tag{10}$$

We choose base learning rate as low as $\gamma_0 = 3{\times}10^{-4}$ and momentum 0.5 in order to achieve convergence. The learning rate is reduced by a factor of 10 throughout training every time the loss reaches a "plateau". More details regarding the learning policy for each experiment are provided in Sec. 3.

Initialization. Initially, the network weights and bias were randomly initialized using the "Xavier" algorithm [17] which automatically determines the variance of initialization for each layer based on the number of input and output neurons. Since learning the parameters of such a large model demands a significant amount of data, we found that multimodal data augmentation can significantly improve the performance. To this end, we initialized parameters θ_g with a pre-trained RGB-based person re-identification model that has been trained on the union of several ReID datasets (*JSTL-DGD* model from [63]). In that case, only the parameters of the added fully-connected layer for training the encoder are initialized at random. The learning rate multipliers for the learnable parameters of that layer are set 10 times larger than all multipliers for the rest of the network.

Regularization. Given the data sparsity, regularizing the model weights is very important for identifying discriminative regions in depth images for person re-identification. We explore two different methods. First, we use the original model without the regularizer. Next, we introduce dropout between the two fully-connected layers ("fc7 dropout"), where most parameters are concentrated. In Fig. 4, we show the benefits of adding noise between layers "fc7" and "fc8", both in terms of top-1 accuracy and generalization ability.

2.3.2 Training the attention unit and the RNN. The parameters of our model $\theta = \{\theta_g, \theta_w, \theta_h, \theta_c\}$ are learned so that the agent maximizes its total reward over time $R = \sum_{t=1}^{T} r_t$ under the distribution of all possible sequences $p(s_{1:T}; \theta)$. This involves calculating the expectation $J(\theta) = \mathbb{E}_{p(s_{1:T};\theta)}[R]$ over a very big number of sequences, which can quickly become intractable. As proposed by Williams [62] and recently deployed successfully on recurrent models of spatial visual attention [20, 43], a sample approximation of the gradient, known as the REINFORCE rule, can be applied as follows:

$$\nabla_\theta J = \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T};\theta)}[\nabla_\theta \log \pi(u_t | s_{1:t}; \theta)(R_t - b_t)] \tag{11}$$

$$\approx \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \nabla_\theta \log \pi(u_t^i | s_{1:t}^i; \theta)(R_t^i - b_t) \tag{12}$$

where s^i's sequences are obtained after M episodes.

In our case, as REINFORCE is applied on the output of Bernoulli stochastic unit with $p = f_w(g_t; \theta_w)$ and probability mass function $\log f(u; p) = u \log p + (1-u)\log(1-p)$, the gradient approximation is given by:

$$\nabla_\theta J \approx \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \frac{u_t^i - p_t^i}{p_t^i(1-p_t^i)}(R_t^i - b_t) \tag{13}$$

where $R_t^i = \sum_{t'=1}^{t} r_{t'}^i$ is the cumulative reward obtained following the execution of action u_t^i. Please note that this is a *biased* estimate of the gradient in order to achieve lower variance, as a baseline reward b_t is used. Consistent with [43], we set $b_t = \mathbb{E}_\pi[R_t]$, which is computed as the mean square error between R_t^i and b_t is minimized by backpropagation. This way, the baseline reward is learned at the same rate as the rest of the model.

All in all, a *hybrid* supervised loss is used to train the attention unit and the RNN's classification output. At each step, the agent takes an action w_t and the reward signal R_t^i is the supervision for evaluating the value of the action for the classification task. The REINFORCE rule increases the log-probability of an action that results in higher accumulated reward than the expected (baseline) total reward (i.e. by increasing $f_w(g_t; \theta_w)$). Otherwise, the log-probability decreases. Finally, in order to backpropagate the gradients through the classifier that is attached on the LSTM unit backward through the whole network, we minimize the cross-entropy loss as is customary in supervised learning. The objective is to maximize the conditional probability of the true label given the observations I_t, i.e. we maximize $\log \pi(c_t^* | s_{1:t}; \theta)$, where c_t^* corresponds to the ground-truth class for time step t.

3 EXPERIMENTS

3.1 Depth-based Datasets

TUM-GAID. Most existing depth-based datasets for person re-identification contain a small number of subjects. *IIT PAVIS* [4], *BIWI* [45] and *IAS-Lab* [46] contain 79, 50 and 11 persons, respectively. On the other hand, *TUM-GAID* database [22] is one of the largest to date. It contains RGB video, depth and audio for 305 people in three variations. A subset of 32 people is recorded a second time after three months with different clothes. Cropped versions for both the RGB and the depth image sequences are provided by the authors. The skeleton data is not available for this dataset.

FaceBody. In spite of its large number of subjects, the persons in TUM-GAID always appear from the side view with fixed viewpoint and distance. As we want to explore the robustness of our method with varying vantage point, human pose, scale, and partial occlusions, we introduce a new dataset, which we name *FaceBody*. It includes 57 subjects appearing from 2 camera viewpoints walking

Dataset	IDs	Training Images	Testing Images	Appearances
TUM-GAID [22] (N-train)	150	34,881	17,625	2 (for 16 IDs)
TUM-GAID [22] (N-test)	155	35,454	17,776	2 (for 16 IDs)
FaceBody (t1, t4)	57	18,178	14,984	2
DPI-T [20]	12	3,740	4,010	5

Table 1: Statistics of the datasets.

into a meeting room in different walking patterns. Each person executes 6 walking sequences, amounting to 12 sequences in total. We simultaneously collect the color and depth sequences with the Kinect V2 Sensor. For a subset of the data sequences where skeleton tracking is successful by Kinect SDK (e.g. when the face is visible at some point or when there are no large body occlusions), we also have the skeleton information, which is the 3D location of 25 human joints and pixel-wise *body index* per person.

Depth-based Person Identification from Top (DPI-T). Haque et al. [20] recently introduced DPI-T for person re-identification from depth. The new dataset contains 12 persons in 300 training and 355 testing sequences for a total of $3,740$ training and $4,010$ testing images, respectively. It is different from previous datasets in many ways. First, more diverse observations per individual are included, as the subjects appear in a total of 25 sequences across many days. The individuals wear 5 different set of clothes on average and walk at variable speeds. Second, unlike most publicly available datasets, the subjects appear from the top. This is a common scenario in automated video surveillance, where the camera is attached near the ceiling looking down. Third, the individuals are captured in daily life situations where they hold objects such as handbags, laptops and coffee. This data imposes new challenges in person re-identification and is used as the third benchmark. Table 1 provides a summary of statistics for the three datasets.

3.2 Evaluation Metrics

Top-k accuracy equals the percentage of test images or sequences for which the ground-truth label is contained within the first k model predictions. Plotting the top-k accuracy as a function of k gives the Cumulative Matching Curve (CMC). Integrating the area under the CMC curve and normalizing for the number of IDs produces the normalized Area Under the Curve (nAUC).

We evaluate our method in both “single-shot" and “multi-shot" mode by testing on individual images and sequences, respectively.

3.3 Experimental Settings

The encoder f_{CNN} is trained using Caffe [25]. Based on the input size of the deployed convolutional architecture, we rescale the input depth images to be 144×56 and subtract the mean depth image. We train our model using stochastic gradient descent with mini-batches of 50 images for training and testing. We set the momentum as low as 0.5, as higher values cause the model to diverge. The momentum μ effectively multiplies the size of the updates by a factor of $\frac{1}{1-\mu}$ after several iterations, so lower values result in smaller updates. The weight decay is set $2 * 10^{-4}$, as is common in Inception type of architecture [57].

The rest of the model in Fig. 3 is implemented in Torch/Lua. We implemented our own customized conversion scripts from Caffe to Torch for the pretrained encoder, as the architecture is not standard. As for training Depth ReID, the batch size is 50 images, the momentum is 0.9 and the learning rate linearly decreases from 0.01 to 0.00001 in 400 epochs up to 500 epochs maximum duration. For the RNN history of $rho = 3$ frames is used, unless otherwise stated.

The experiments are conducted on a modern machine with NVIDIA Tesla $K80$ GPU, 24 Intel Xeon E5 cores and $64G$ RAM memory. The code implementing our method and the pretrained models necessary to reproduce the evaluation will be distributed publicly upon completion of the anonymous review process.

3.4 Baselines

Color model. The model designed by Xiao et al. [63] has been shown to outperform other methods on various public datasets. For instance, they achieve 13.2% higher CMC top-1 accuracy than the previous top-performing method [48] on large CUHK03 [31]. Therefore, we choose this method as our RGB-based baseline.

Motion model. We also compare our method to a motion-based method, as motion is also insensitive to appearance changes. Castro et al. [7] demonstrated competitive results on TUM-GAID, although they used a resolution of 80×60, which is 8 times lower than the original resolution of 640×480 for these sequences. By comparison, our model's input is 144×56. Additionally, although we can make one-shot predictions, Castro et al. built a representation on subsequences of 25 frames. They extract dense optical flow between consecutive frames, crop and stack the flow channels, which are then passed through a convolutional neural network to obtain gait signatures for the entire subsequence.

Depth model. The Recurrent Attention Model (RAM) introduced by Haque et al. [20] relies only on depth images like our method. They introduced the DPI-T dataset, which we use for comparisons.

3.5 TUM-GAID database

Evaluation protocol. TUM-GAID depth data includes 12 “normal" sequences (N), 4 sequences with a backpack (B) and 4 sequences with coating shoes (S). We use the N setting, where sequences $n01$–$n06$ are from session 1, and sequences $n07$–$n12$ are from session 2, where the subjects have changed clothes. In half of the sequences the persons walk from left to right, while in the other half they walk from right to left. Of the 305 persons that appear in session 1, only 32 of them participate in session 2. Based on the official protocol, we use sequences $n1$–$n4$, $n07$–$n10$ for training, and sequences $n5$–$n6$ and $n11$–$n12$ for validation and testing, respectively. The subjects are partitioned into 150 training and 155 testing subjects, where the split is even for individuals participating in session 2.

Preprocessing. The tracked RGB and depth data are conveniently provided by the creators of TUM-GAID. Since the skeleton data are not available, we do not perform background removal. This has minor influence, as the background is identical for all sequences, filled in with a plain wall.

Resolution	Method	top-1 accuracy (%)
640 × 480	*Gait Energy* [22]	44.0
	SVIM [61]	65.6
	Fisher Motion [8]	78.1
	SDL [71]	96.9
80 × 60	*Gait Signatures* [7]	62.5
144 × 56	*RGB ReID* (TL) [63]	70.5
	RGB ReID [63]	94.7
	Depth ReID (TL)	92.7
	Binary Depth ReID	95.4
	Depth ReID	**97.0**

Table 2: Comparisons on TUM-GAID for Task 1.

Task 1: Training on multiple clothes. First, we use all training sequences where the individuals appear in two sets of clothes. For this experiment we exclusively benchmark the f_{CNN} module. It is pre-trained on the training subjects, and afterwards fine-tuned on the training sequences of the testing subjects. Small base learning rate of 5×10^{-4} is used for pre-training. For fine-tuning the base rate is set 1×10^{-3}, as the network has adapted to depth data. A multistep policy is adopted where the learning rate decreases by a factor of 10 after $8k$ and $12k$ iterations and the training converges by $16k$ iterations. Since the Color ReID network [63] is already pre-trained on RGB-based datasets, we directly fine-tune it on the testing subjects. Finally, we train a depth model with the same protocol on binary representations $\mathbf{D}_\mathbf{p}^\mathbf{b}$ (see Sec. 2.1).

In Table 2 we provide comparison with other methods. Since the motion-based baseline [7], which also uses a deep convolutional architecture, allows fine-tuning only the top layer on the testing IDs, we also evaluate Color and Depth ReID with this constraint. These methods are presented at rows 6 and 8 notated as "TL". The deployed resolution that different methods use is noteworthy. Most methods under comparison use the data in their original resolution, which is 640×480. Our method and the other two methods that are based on convolutional networks downsample the images by a large factor in order to match the model input. Despite its lower resolution, Depth ReID outperforms the other methods. Additionally, even when fine-tuning only the last layer, the depth features are well-transferable [69] to the new set of persons.

Task 2: Constrained training on one set of clothes. Our objective is to examine whether a color-insensitive representation such as depth can offer more accurate person re-identification when the subjects change clothes. To that end, we fine-tune on the training sequences $n01$–$n04$ of the testing IDs, using the sequences $n05$–$n06$ for validation. Therefore, this model has no access to training data from session 2. Next, the model is evaluated on sequences $n11$–$n12$. We make the assumption that the 32 subjects that participate in the second recording are known.

In Table 3 we show that Depth ReID is more robust than the corresponding RGB model, presenting 6.2% higher top-1 and 23.6% higher top-5 accuracy in single-shot mode. Note that Depth ReID achieves 97.0% accuracy (cf. Table 2) when sequences from both set of clothes are available during training. This is a critical problem to deal with as training data are not always available for new clothes.

Method	top-1	top-5	nAUC
RGB ReID, single-shot [63]	41.8	64.4	74.3
Depth ReID, single-shot	48.0	88.0	85.0
Depth, single-shot+RGB ReID	48.6	83.0	81.9
Head RGB ReID	59.2	78.4	79.4
Depth, single-shot+Head RGB ReID	65.4	85.9	85.2
Depth ReID, multi-shot with RNN	56.3	87.5	87.5
Depth ReID, multi-shot with RNN and attention	59.4	93.8	89.6
Head RGB ReID+Depth ReID, multi-shot with RNN and attention	71.9	93.8	89.9

Table 3: Recognition accuracy (%) and normalized area under the curve (%) on TUM-GAID (normal sequences) for Task 2.

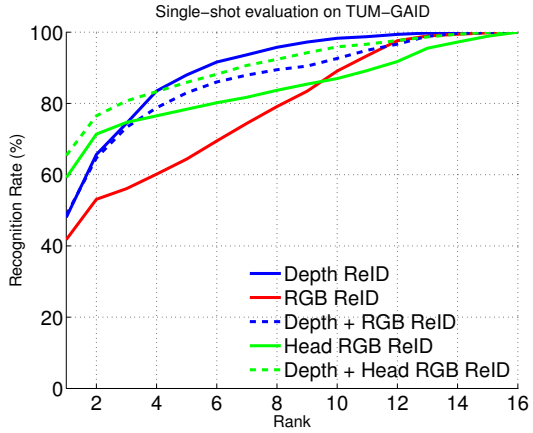


Figure 5: Cumulative matching curves for Task 2 on TUM-GAID. For rank-k (x axis), the y axis denotes recognition accuracy, if the ground truth label is within the method's top-k predictions.

As large variations in color and texture can be distracting for verification purposes, we attempt to rely more on the head region, which is less sensitive to day-by-day changes. To that purpose, we fine-tune the RGB-based pre-trained model [63] on the upper body part, which we call "Head RGB ReID". In order to remove the foreground, we extract a binary mask from depth by thresholding the depth representation. Given that the subjects in color and depth images are not perfectly pixel aligned, we apply morphological dilation on the binary mask with a circular disk of radius 4 to ensure that the foreground region includes the whole body in RGB. Then we crop the head region using the circumscribed rectangle around the top 1/4 of the foreground region. In Table 3 we see the improvement in top-1 accuracy using Head RGB ReID, individually and jointly with depth information. Finally, we show the accuracy of Depth ReID with LSTM units and temporal attention, while evaluating each sequence in multi-shot mode.

In Fig. 5 we visualize the CMC curves for single-shot setting. Depth ReID scales better than its counterparts, which is validated by

Method	top-1	top-5
RGB ReID, single-shot [63]	62.7	90.6
Depth ReID, single-shot	78.6	91.4
Depth ReID, Multi-shot with RNN	91.1	98.3
Depth ReID, Multi-shot with RNN and attention	92.9	98.8

Table 4: Re-identification accuracy (%) on FaceBody.

Setting	Method	top-1	top-5
Single-shot	*3D RAM* [20]	47.5	—
	Depth ReID, single-shot	62.3	93.6
Multi-shot	*4D RAM* [20]	55.6	—
	Depth ReID, averaging	72.6	96.4
	Depth ReID, RNN with Bernoulli p=0.5	73.9	96.4
	Depth ReID, RNN with learned Bernoulli p	75.9	96.0
	Depth ReID, RNN and attention	77.5	96.0

Table 5: Re-identification accuracy (%) on DPI-T [20].

the normalized Area Under the Curve (nAUC) in Table 3. Intuitively, when the face is well-visible, Head RGB ReID is expected to be reliable, which explains the higher top-1 accuracy. On the other hand, when the face is mostly occluded, more guesses are not likely to improve the re-identification rate more quickly than body models.

3.6 FaceBody dataset

Evaluation protocol. The new dataset contains 6 sequences, $t1$–$t6$, of 57 subjects in a realistic meeting room scenario, as captured by two different viewpoints with Kinect V2 Sensors. The persons enter the room, walk in various paths, write on the board, and then exit the room. The data, in addition to RGB and depth images, includes the skeleton tracking, i.e. the body index information, which is pixel aligned to the depth images and the $3D$ location of 25 pre-determined joints [53]. The body index is a reliable way to crop the persons in all frames, while sparing the need to deploy a tracker. However, the body index is available only when the skeleton tracking works successfully. In order to ensure the quality of extracted detections, we use the sequences $t1$ and $t4$ from each camera that have skeleton data for all 57 subjects. Let us denote the two cameras $c1$ and $c2$. The sequences $t1/c1$ and $t1/c2$ are used for training and the sequences $t4/c1$ and $t4/c2$ for testing, which sum up to 18, 178 training and 14, 984 validation images.

Preprocessing. We follow the process as described in Sec. 2.1 to obtain the depth crops $\mathbf{D}_{\mathbf{p}}^{\mathbf{g}}$. As there is no perfect alignment between the depth and the RGB data, we do not mask out the background for the RGB images. Therefore, the background is a nuisance for Color ReID on FaceBody. However, all sequences are recorded in the same room, so the background should have limited effect. Instead, for RGB images, we use the body index to extract a rectangular region around the person and add a 20–$pixel$ margin to ensure that the person's silhouette lies within the bounding box.

Comparisons (Table 4). Although FaceBody poses new challenges, as the subjects present pose variation and partial occlusions, Depth ReID is consistently more reliable than Color ReID. Part of this improvement can be attributed to the precise background subtraction based on body index in case of depth, which yields very accurate global shape information.

Multi-shot evaluation. Following, we leverage on multiple frames from each sequence to make k-shot predictions. In order to make the evaluation more challenging, we allow to use only $k = 3$ consecutive frames per evaluation. This is a realistic scenario where a tracked person can be occasionally occluded or there is lack of motion. Most testing sequences have length N in the order of 200 frames. Our protocol is to perform 100 runs for each sequence where the start frame is chosen uniformly at random in $\{1, \ldots, N - k + 1\}$ range. In Table 4 we show the performance of Depth ReID with LSTM units and temporal attention.

Inspecting the temporal attention unit. After inspecting the estimated Bernoulli parameter $p = f_w(g_t; \theta_w)$ on unknown testing data, we observed that large variations are possible within one sequence, even between neighboring frames. Lower values are usually associated with noisy frames as in the example sequence in Fig. 6, or with challenging human pose and partial occlusions which are not well represented in the training set.

3.7 DPI-T dataset

Depth ReID is trained on DPI-T following the procedure described in Sec. 2.3 and the official evaluation protocol. For the multi-shot setting all available frames are used for a single sequence prediction. Although the persons on DPI-T have many more appearances (5 different sets of clothes on average), the sequences are shorter than in the other two datasets (approximately 16 frames per sequence).

In Table 5 we demonstrate our model's performance compared to Haque et al. [20]. For single-frame predictions we use only the encoder f_{CNN} with its attached classifier. For multi-shot mode with averaging in row 4, we simply calcuate the average of f_{CNN} outputs over each sequence frame. Next in rows $5 - 7$ we show results with LSTM units attached on the encoder. As for the last row, each sequence's class posterior is computed as the weighted sum of the model's outputs c_t for the sequence length K, based on the inferred weights $w_1, \ldots, w_K$. In rows 5 and 6 all frames contribute equally. Note that the RNN with constant Bernoulli $p = 0.5$ performs worse than the model which learns p. It is to be expected that learning the parameter p via the attention unit enforces learning better models, as frames are preserved or dropped out based on how likely they are to increase the accumulated reward and not uniformly at random such as when $p = 0.5$.

4 DISCUSSION

We have presented a novel framework for person re-identification in the absence of RGB information, hence in the dark. Our pipeline

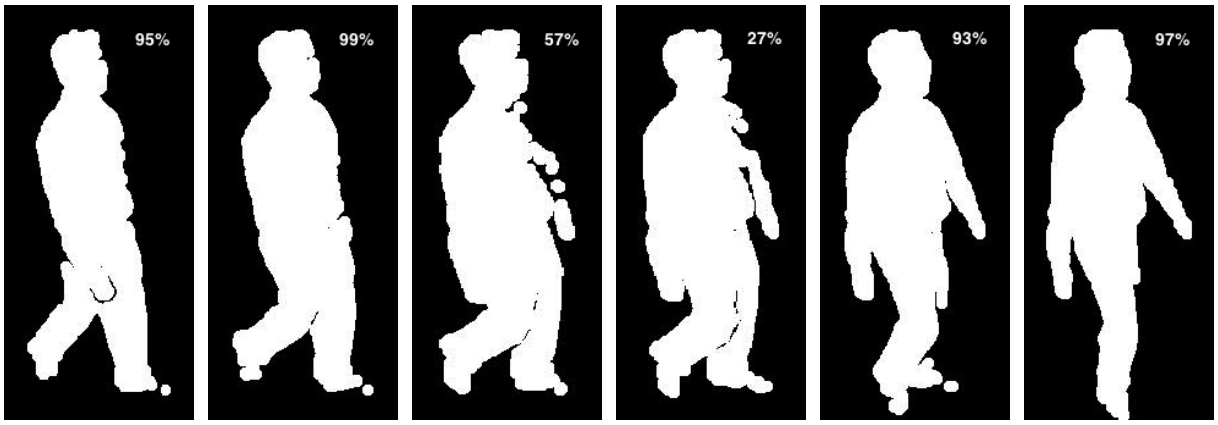


Figure 6: Example sequence along with the inferred Bernoulli parameter $p = f_w(g_t; \theta_w) \in [0, 100](\%)$ using a trained Depth ReID model with attention on FaceBody. Frames that are characterized by noisy measurements, uncommon pose and partial occlusions are likely to contribute less in multi-shot prediction, based on the estimated weight by the temporal attention unit.

leverages grayscale encodings from depth measurements, normalized, offset and masked using skeleton information and morphology, in order to learn depth representations with a recurrent deep convolutional architecture. We tackle the small sample size problem with regularizers and by introducing a temporal attention unit that allows efficient and scalable training with video sequences. The entire model can be trained end to end with a hybrid supervised loss with principles to maximize the conditional probability of the true class identity and the REINFORCE rule. Although not necessary in our pipeline, note that the model can be extended to calculate spatio-temporal attention regions.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3908–3916.

[2] Antonio Albiol, J Oliver, and JM Mossi. 2012. Who is who at different cameras: people re-identification using depth cameras. *IET computer vision* 6, 5 (2012), 378–387.

[3] Virginia Andersson, Rafael Dutra, and Ricardo Araújo. 2014. Anthropometric and human gait identification using skeleton data from kinect sensor. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. ACM, 60–61.

[4] Igor Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. 2012. Re-identification with rgb-d sensors. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 433–442.

[5] Apurva Bedagkar-Gala and Shishir K Shah. 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing* 32, 4 (2014), 270–286.

[6] Léon Bottou. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. Springer, 421–436.

[7] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, and N Pérez de la Blanca. 2016. Automatic learning of gait signatures for people identification. *arXiv preprint arXiv:1603.01006* (2016).

[8] Francisco M Castro, Manuel J Marín-Jimenez, and Rafael Medina-Carnicer. 2014. Pyramidal fisher motion for multiview gait recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 1692–1697.

[9] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211* (2015).

[10] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. 2016. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1268–1277.

[11] Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu. 2000. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern recognition* 33, 10 (2000), 1713–1726.

[12] Yeong-Jun Cho and Kuk-Jin Yoon. 2016. Improving person re-identification via pose-aware multi-shot matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1354–1362.

[13] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48, 10 (2015), 2993–3003.

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[15] Amandine Dubois and François Charpillet. 2014. A gait analysis method based on a depth camera for fall prevention. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 4515–4518.

[16] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2360–2367.

[17] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks.. In *Aistats*, Vol. 9. 249–256.

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

[19] Douglas Gray and Hai Tao. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Computer Vision–ECCV 2008* (2008), 262–275.

[20] Albert Haque, Alexandre Alahi, and Li Fei-Fei. 2016. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1229–1238.

[21] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[22] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. 2014. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation* 25, 1 (2014), 195–206.

[23] Dimosthenis Ioannidis, Dimitrios Tzovaras, Ioannis G Damousis, Savvas Argyropoulos, and Konstantinos Moustakas. 2007. Gait recognition using compact feature extraction transforms and depth information. *IEEE Transactions on Information Forensics and security* 2, 3 (2007), 623–630.

[24] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

[26] Amir Kale, Naresh Cuntoor, B Yegnanarayana, AN Rajagopalan, and Rama Chellappa. 2003. Gait analysis for human identification. In *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 706–714.

[27] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. 2016. Person Re-Identification by Unsupervised\ ell _1 Graph Learning. In *European Conference on Computer Vision*. Springer, 178–195.

[28] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2288–2295.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[30] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. 2013. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 7 (2013), 1622–1634.

[31] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[32] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. 2013. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3610–3617.

[33] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2197–2206.

[34] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. 2015. Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence* 37, 8 (2015), 1629–1642.

[35] Giuseppe Lisanti, Iacopo Masi, and Alberto Del Bimbo. 2014. Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*. ACM, 10.

[36] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[37] Bingpeng Ma, Yu Su, and Frédéric Jurie. 2012. Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 413–422.

[38] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. 2014. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing* 23, 8 (2014), 3656–3670.

[39] Al Mansur, Yasushi Makihara, Rasyid Aqmar, and Yasushi Yagi. 2014. Gait recognition under speed transition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2521–2528.

[40] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. 2016. Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*. Springer, 858–877.

[41] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. 2016. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1325–1334.

[42] Alexis Mignon and Frédéric Jurie. 2012. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2666–2672.

[43] Volodymyr Mnih, Nicolas Heess, Alex Graves, and others. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*. 2204–2212.

[44] A Mogelmose, Thomas B Moeslund, and Kamal Nasrollahi. 2013. Multimodal person re-identification using RGB-D sensors and a transient identification database. In *Biometrics and Forensics (IWBF), 2013 International Workshop on*. IEEE, 1–4.

[45] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 2014. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 4512–4519.

[46] Matteo Munaro, Stefano Ghidoni, Deniz Tartaro Dizmen, and Emanuele Menegatti. 2014. A feature-based approach to people re-identification using skeleton keypoints. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 5644–5651.

[47] Brent Munsell, Andrew Temlyakov, Chengzheng Qu, and Song Wang. 2012. Person identification using full-body motion and anthropometric biometrics from kinect videos. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 91–100.

[48] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. 2015. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1846–1855.

[49] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3318–3325.

[50] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. 2010. Person Re-Identification by Support Vector Ranking.. In *BMVC*, Vol. 2. 6.

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[52] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. 2016. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*. Springer, 732–748.

[53] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013), 116–124.

[54] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[55] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2016. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*. Springer, 475–491.

[56] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1891–1898.

[57] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[58] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. 2013. Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology* 23, 10 (2013), 1675–1685.

[59] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. 2013. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)* 46, 2 (2013), 29.

[60] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1288–1296.

[61] Tenika Whytock, Alexander Belyaev, and Neil M Robertson. 2014. Dynamic distance-based shape features for gait recognition. *Journal of Mathematical Imaging and Vision* 50, 3 (2014), 314–326.

[62] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[63] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1249–1258.

[64] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. 2014. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*. Springer, 1–16.

[65] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. 2016. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*. Springer, 701–716.

[66] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. 2014. Salient color names for person re-identification. In *European Conference on Computer Vision*. Springer, 536–551.

[67] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. 4507–4515.

[68] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 34–39.

[69] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[70] Wojciech Zaremba and Ilya Sutskever. 2014. Learning to execute. *arXiv preprint arXiv:1410.4615* (2014).

[71] Wei Zeng, Cong Wang, and Feifei Yang. 2014. Silhouette-based gait recognition via deterministic learning. *Pattern recognition* 47, 11 (2014), 3568–3584.

[72] Li Zhang, Tao Xiang, and Shaogang Gong. 2016. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1239–1248.

[73] Guoying Zhao, Guoyi Liu, Hua Li, and Matti Pietikainen. 2006. 3D gait recognition using multiple cameras. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 529–534.

[74] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 2528–2535.

[75] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3586–3593.
[76] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2014. Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 144–151.
[77] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*. Springer, 868–884.
[78] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.
[79] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2013. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence* 35, 3 (2013), 653–668.