Domain Adaptation Through Synthesis for Unsupervised Person Re-identification

Sławomir Bąk[1](✉), Peter Carr[1], and Jean-François Lalonde[2]

[1] Argo AI, Pittsburgh, PA 15222, USA
{sbak,pcarr}@argo.ai
[2] Université Laval, Quebec City G1V 0A6, Canada
jflalonde@gel.ulaval.ca

Abstract. Drastic variations in illumination across surveillance cameras make the person re-identification problem extremely challenging. Current large scale re-identification datasets have a significant number of training subjects, but lack diversity in lighting conditions. As a result, a trained model requires fine-tuning to become effective under an unseen illumination condition. To alleviate this problem, we introduce a new synthetic dataset that contains hundreds of illumination conditions. Specifically, we use 100 virtual humans illuminated with multiple HDR environment maps which accurately model realistic indoor and outdoor lighting. To achieve better accuracy in unseen illumination conditions we propose a novel domain adaptation technique that takes advantage of our synthetic data and performs fine-tuning in a completely unsupervised way. Our approach yields significantly higher accuracy than semi-supervised and unsupervised state-of-the-art methods, and is very competitive with supervised techniques.

Keywords: Synthetic · Identification · Unsupervised
Domain adaptation

1 Introduction

Even over the course of just a few minutes, a person can look surprisingly different when observed by different cameras at different locations. Indeed, her visual appearance can vary drastically due to changes in her pose, to the different illumination conditions, and to the camera configurations and viewing angles. To further complicate things, she may be wearing the same shirt as another, unrelated person, and could thus easily be confused.

The task of person re-identification tackles the challenge of finding the same subject across a network of non-overlapping cameras. Most effective state-of-the-art algorithms employ supervised learning [25–27,46,51], and require thousands

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01261-8_12) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11217, pp. 193–209, 2018.
https://doi.org/10.1007/978-3-030-01261-8_12

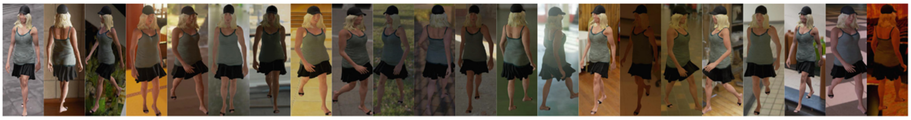

Fig. 1. Sample images from our **SyRI** dataset: the same 3D character rendered in various HDR environment maps. The dataset provides 100 virtual humans rendered in 140 realistic illumination conditions.

of labeled images for training. With novel deep architectures, we are witnessing an exponential growth of large scale re-identification datasets [25,48]. Recent re-identification benchmarks have focused on capturing large numbers of identities, which allows the models to increase their discriminative capabilities [43]. Unfortunately, current re-identification datasets lack significant diversity in the number of lighting conditions, since they are usually limited to a relatively small number of cameras (the same person is registered under a handful of illumination conditions). Models trained on these datasets are thus biased to the illumination conditions seen during training. One can increase the model generalization by merging multiple re-identification datasets into a single dataset and training the network as joint single-task learning [43]. In this approach, the learned models show generalization properties but only upon fine-tuning [2]. This is because the merged datasets contain tens of different lighting conditions, which might not be sufficient to generalize. To apply the previously trained model to a new set of cameras, we need to annotate hundreds of subjects in each camera, which is a tedious process and does not scale to real-world scenarios.

In this work, we introduce the ***Sy****nthetic Person* ***R****e*-***I****dentification* (**SyRI**) dataset. Employing a game engine, we simulate the appearance of hundreds of subjects under different realistic illumination conditions, including indoor and outdoor lighting (see Fig. 1). We first carefully designed 100 virtual humans based on 3D scans of real people. These digital humans are then rendered using realistic backgrounds and lighting conditions captured in a variety of high dynamic range (HDR) environment maps. We use HDR maps as the virtual light source and background plate when rendering the 3D virtual scenes. With the increased diversity in lighting conditions, the learned re-identification models gain additional generalization properties, thus performing significantly better in unseen lighting conditions.

To further improve recognition performance, we propose a novel **three-step domain adaptation technique**, which translates our dataset to the target conditions by employing cycle-consistent adversarial networks [52]. Since the cycle-consistent formulation often produces semantic shifts (the color of clothing may change drastically during translation), we propose an additional regularization term to limit the magnitude of the translation [37], as well as an additional masking technique to force the network to focus on the foreground object. The translated images are then used to fine-tune the model to the specific lighting conditions. In summary, our main contributions are:

- We introduce a new dataset with 100 virtual humans rendered with 140 HDR environment maps. We demonstrate how this dataset can increase generalization capabilities of trained models in unseen illumination conditions without fine-tuning.
- We improve re-identification accuracy in an unsupervised fashion using a novel three-step domain adaptation technique. We use cycle-consistency translation with a new regularization term for preserving identities. The translated synthetic images are used to fine-tune the re-identification model for a specific target domain.

2 Related Work

Person Re-identification: Most successful person re-identification approaches employ supervised learning [3,22,23]. This includes novel deep architectures and the debate as to whether the triplet or multi-classification loss is more effective for training re-identification networks [3,16,43]. Larger architectures have improved accuracy, but also increased the demand for larger re-identification datasets [12,47,48]. However all of these approaches require fine-tuning [2,47] to become effective in unseen target illumination conditions, which is infeasible for large camera networks. To overcome this scalability issue, semi-supervised and unsupervised methods have been proposed [20,21,35]. This includes transfer learning [18,35,49] and dictionary learning [1,10,28]. However, without labeled data, these techniques usually look for feature invariance, which reduces discriminativity, and makes the methods uncompetitive with supervised techniques.

Synthetic Data: Recently, data synthesis and its application for training deep neural architectures has drawn increasing attention [37]. It can potentially generate unlimited labeled data. Many computer vision tasks have already been successfully tackled with synthetic data: human pose estimation [36], pedestrian detectors [7,14,19] and semantic segmentation [30,34]. The underlying challenge when training with synthetic visual data is to overcome the significant differences between synthetic and real image statistics. With increasing capacity of neural networks, there is a risk that the network will learn details only present in synthetic data and fail to generalize to real images. One solution is to focus on rendering techniques to make synthetic images appear more realistic. However, as the best renderers are not differentiable, the loss from the classifier cannot be directly back-propagated, thus leaving us with simple sampling strategies [19]. Instead, we take an approach closer to [37]: rather than optimizing renderer parameters, we cast the problem as a domain adaptation task. In our case, the domain adaptation performs two tasks simultaneously: (1) makes the synthetic images look more realistic and (2) minimizes the domain shift between the source and the target illumination conditions.

Domain Adaptation: Typically, domain adaptation is a way of handling dataset bias [40]. Not surprisingly, domain adaptation is also used to minimize the visual gap betwen synthetic and real images [37]. Often this shift between distributions of the source and target domain is measured by the distance between

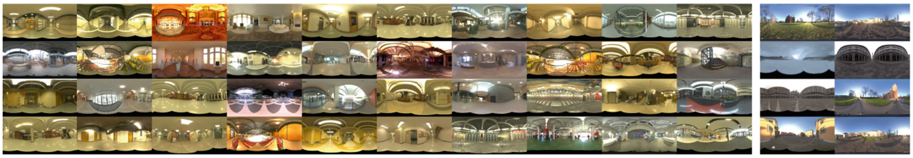

Fig. 2. Example HDR environment maps used to relight virtual humans. The environment maps capture a wide variety of realistic indoor (left) and outdoor (right) lighting conditions. The images have been tonemapped for display purposes with $\gamma = 2.2$. Please zoom-in for more details.

the source and target subspace representations [8]. Thus, many techniques focus on learning feature space transformations to align the source and the target domains [18,41]. This enables knowledge transfer (*e.g.* how to perform a particular task) between the two domains. Recently, adversarial training has achieved impressive results not only in image generation [11], but also in unsupervised domain adaptation [9]. In this work, we are inspired by a recent approach for unsupervised image-to-image translation [52], where the main goal is to learn the mapping between images, rather than maximizing the performance of the model in particular task. Given our synthesized images and the domain translation, we are able to hallucinate labeled training data in the target domain that can be used for fine-tuning (adaptation).

3 SyRI Dataset

Given sufficient real data covering all possible illumination variations, we should be able to learn re-identification models that have good generalization capabilities without the need for fine-tuning. Unfortunately, gathering and annotating such a dataset is prohibitive. Instead, we propose training with synthesized data. The underlying challenge is to create photo-realistic scenes with realistic lighting conditions. Rather than hand-crafting the illumination conditions, we use High Dynamic Range (HDR) environment maps [5]. These can be seen as 360° panoramas of the real world that contain accurate lighting information, and can be used to relight virtual objects and provide realistic backgrounds.

Environment Maps. To accurately model realistic lighting, a database of 140 HDR environment maps was acquired. First, 40 of those environment maps were gathered from several sources online[1]. Further, we also captured an additional 100 environment maps. A Canon 5D Mark III camera with a Sigma 8 mm fisheye lens was mounted on a tripod equipped with panoramic tripod head. 7 bracketed exposures were shot at 60° increments, for a total of 42 RAW photos per

[1] The following online sources were used: http://gl.ict.usc.edu/Data/HighResProbes/, http://dativ.at/lightprobes, http://www.unparent.com/photos_probes.html, http://www.hdrlabs.com/sibl/archive.html.

panorama. The resulting set of photos were automatically merged and stitched into a 22 f-stop HDR 360° environment map using the PTGui Pro commercial software. Our dataset represents a wide variety of indoor and outdoor environments, such as offices, theaters, shopping malls, museums, classrooms, hallways, corridors, etc. Figure 2 shows example environment maps from our dataset.

3D Virtual Humans and Animations. Our 3D virtual humans are carefully designed with *Adobe Fuse CC* that provides 3D content, including body scans of real people with customizable body parts and clothing. We generate 100 character prototypes, where we customize body shapes, clothing, material textures and colors (see Fig. 3). These characters are then animated using rigs to obtain realistic looking walking poses.

Rendering. We use *Unreal Engine 4* to achieve real-time rendering speeds. To relight our 3D virtual humans, the HDR environment map is texture mapped on a large sphere surrounding the scene. This sphere is then used as a the sole light source (light emitter) to render the scene. We position a 3D character at the center of the sphere. The character is animated using either a male or female walking rig, depending on the model gender. We also add a rotation animation to acquire multiple viewpoints of each subject. The camera position is matched with existing re-identification datasets. Each subject is rendered twice under the same HDR map rotating the sphere about its vertical axis by two random angles. This effectively provides two different backgrounds and lighting conditions for each environment map. We render 2-second videos at 30 fps as the character is being rotated. In the end, we render 100 (subjects) × 140 (environment maps) × 2 (rotations) × 2 (seconds) × 30 (fps) $= 1,680,000$ frames. Both the rendered dataset as well as the Unreal Engine project that will allow a user to render more data are going to be made publicly available.

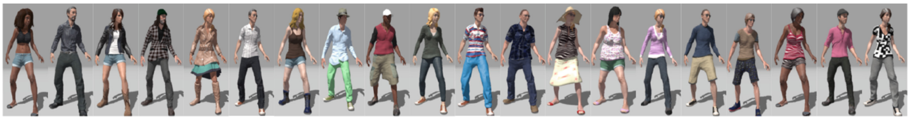

Fig. 3. Sample 3D virtual humans from **SyRI** dataset.

4 Method

We cast person re-identification as a domain adaptation problem, where the domain is assumed to be an illumination condition (*i.e.*, a camera-specific lighting). Our objective is to find an effective and unsupervised strategy for performing person re-identification under the target illumination condition.

For training, we assume we have access to M real source domains $\mathbf{R} = \{R_1 \ldots R_M\}$, where each $R_m = \{x_i, y_i\}_{i=1}^{Z_{R_m}}$ consists of Z_{R_m} real images x_i

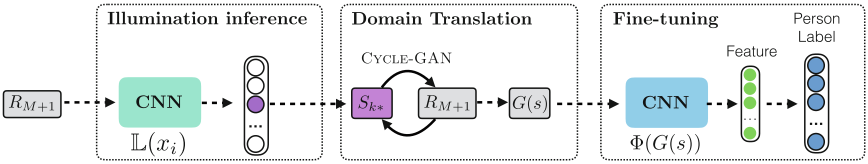

Fig. 4. Unsupervised Domain Adaptation. Given unlabelled input images from target domain R_{M+1}, we first select the closest synthetic domain S_{k*} through illumination inference. Afterwards, images from the selected domain S_{k*} are translated by $G : S_{k*} \rightarrow R_{M+1}$ to better resemble the input images in R_{M+1}. Finally, the translated synthetic images $G(s)$ along with their known identities are used to fine-tune the re-identification network $\varPhi(\cdot)$.

and their labels y_i (person's identity); and N source synthetic domains $\mathbf{S} = \{S_1 \ldots S_N\}$, where each $S_n = \{s_i, y_i\}_{i=1}^{Z_{S_n}}$ consists of Z_{S_n} synthetic images s_i and their labels y_i (3D character's identity). In our case $N \gg M$ as we have access to hundreds of different illumination conditions (see Sect. 3). Our ultimate goal is to perform re-identification in unknown target domain $R_{M+1} = \{x_i\}_{i=1}^{Z_{R_{M+1}}}$ for which we do not have labels.

4.1 Joint Learning of Re-identification Network

We first learn a generic image feature representation for person re-identification. The feature extractor $\varPhi(\cdot)$ is a Convolutional Neural Network (CNN) trained to perform multi-classification task, *i.e.* given a cropped image of a person, the CNN has to predict the person's identity. We propose to merge all domains $\mathbf{R}$ and $\mathbf{S}$ into a single large dataset and train the network jointly from scratch. We adopt the CNN model from [43]. To learn discriminative and generalizable features, the number of classes during training has to be significantly larger than the dimensionality of the last hidden layer (feature layer). In our case the training set consists of 3K+ classes (identities) and the feature layer has been fixed to 256 dimensions.

One could assume that with our new dataset, the pre-trained model should generalize well in novel target conditions. Although synthetic data helps (see Sect. 5.1), there is still a significant performance gap between the pre-trained model and its fine-tuned version on the target domain. We believe there are two reasons for this gap: (1) our dataset does not cover all possible illumination conditions, and (2) there is a gap between synthetic and real image distributions [37]. This motivates the investigation of domain adaptation techniques that can potentially address both issues: making the synthetic images looking more realistic, as well as minimizing the shifts between source and target illumination conditions.

4.2 Domain Adaptation

We formulate domain adaptation as the following three-step process, as illustrated in Fig. 4.

1. **Illumination inference**: find the closest illumination condition (domain $S_{k*} \in \mathbf{S}$) for a given input R_{M+1}.
2. **Domain translation**: translate domain S_{k*} to R_{M+1}, by learning G, $G: S_{k*} \rightarrow R_{M+1}$ while preserving a 3D character's identity from $s \in S_{k*}$.
3. **Fine-tuning**: update $\varPhi(\cdot)$ with the translated domain $G(s)$.

Illumination Inference. Domain adaptation is commonly called a visual dataset bias problem. Dataset bias was compellingly demonstrated in computer vision by the *name the dataset* game of Torralba and Efros [40]. They trained a classifier to predict which dataset an image originated from, illustrating that visual datasets are biased samples of the visual world. In this work, we employ a similar idea to identify the synthetic domain $S_{k*} \in \mathbf{S}$ that is closest to the target domain R_{M+1}. To do so, we train a CNN classifier that takes an input image and predicts which illumination condition the image was rendered with. In our case, the classifier has to classify the image into one of $N = 140$ classes (the number of different environment maps in our synthetic dataset). We used Resnet-18 [15] pretrained on ImageNet and fine-tuned to perform illumination classification. Given the trained classifier, we take a set of test images from R_{M+1} and predict the closest lighting condition by

$$k^* = \underset{k \in \{1 \ldots N\}}{\arg\max} \sum_{i=1}^{Z_{R_{M+1}}} \varDelta\big(\mathbb{L}(x_i), k\big), \quad \mathbf{s.t.} \quad \varDelta\big(\mathbb{L}(x_i), k\big) = \begin{cases} 1, & L(x_i) = k \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

Here, k corresponds to domain class, $\mathbb{L}(x_i)$ is the class predicted by the CNN classifier and $\varDelta$ is a counting function. We use this formulation to find S_{k*}: the synthetic illumination condition that is most similar to the target domain R_{M+1} (*i.e.* requiring the minimum amout of domain shift). S_{k*} will be used to translate images from S_{k*} to R_{M+1} while preserving each 3D character's identity.

Domain Translation. Given two domains S and R (for convenience we skip sub-indices here) and the training samples $s_i \in S$ and $x_i \in R$, our objective is to learn a mapping function $G: S \rightarrow R$. As we do not have corresponding pairs between our synthetic and real domains, G is fairly unconstrained and standard procedures will lead to the well-known problem of mode collapse (all input images map to the same output image). To circumvent this problem, we adapt the technique of [52], where rather than learning a single mapping $G: S \rightarrow R$, we exploit the property that translation should be *cycle-consistent*. In other words there should exist the opposite mapping $F: R \rightarrow S$, where G and F are inverses of each other.

We train both mappings G and F simultaneously, and use two *cycle consistency losses* to regularize the training: $s \rightarrow G(s) \rightarrow F(G(s)) \approx s$ and $x \rightarrow$

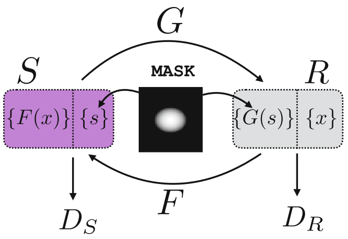

Fig. 5. Semantic Shift Regularization. The Cycle-GAN loss only applies to $F(G(s))$ and $G(F(x))$. There is no constraint on what $G()$ and $F()$ can do individually, which can result in drastic color changes. We incorporate an additional regularization loss requiring s and $G(s)$ to be similar. The loss should only apply to the foreground (to preserve identity), since the target camera may have a very different background than the synthetic data.

$F(x) \rightarrow G(F(x)) \approx x$. G and F are generator functions, where G tries to generate images $G(s)$ that look similar to images from domain R, and F generates images $F(x)$ that should look like images from domain S. Additionally, two adversarial discriminators D_S and D_R are trained, where D_S tries to discriminate between images $\{s\}$ and translated images $\{F(x)\}$; and analogously D_R aims to distinguish between $\{x\}$ and $\{G(s)\}$ (see Fig. 5).

The training objective contains *adversarial losses* [11] for both G and F, as well as two *cycle consistency losses*. The *adversarial loss* for G is defined as

$$\mathcal{L}_{GAN}(G, D_R, S, R) = \mathbb{E}_{x\sim p_{data}(x)}[\log D_R(x)] + \mathbb{E}_{s\sim p_{data}(s)}[\log(1 - D_R(G(s)))], \tag{2}$$

and we can analogously define *adversarial loss* for F, *i.e.* $\mathcal{L}_{GAN}(F, D_S, R, S)$. Both *cycle consistency losses* can be expressed as

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{s\sim p_{data}(s)}[||F(G(s)) - s||_1] + \mathbb{E}_{x\sim p_{data}(x)}[||G(F(x)) - x||_1]. \tag{3}$$

The final objective is

$$\mathcal{L}_{CycleGAN}(G, F, D_S, D_R) = \mathcal{L}_{GAN}(G, D_R, S, R) + \mathcal{L}_{GAN}(F, D_S, R, S) + \lambda_1 \mathcal{L}_{cyc}(G, F), \tag{4}$$

where λ_1 controls the relative importance of the *cycle consistency losses*.

Semantic Shift Regularization. In the above formulation, there is no constraint that the color distribution of the generated image $G(s)$ should be close to instance s. With large capacity models, the approach can map the colors within s to any distribution, as long as this distribution is indistinguishable from the emperical distribution within R ($F(x)$ will learn the inverse mapping). In our application, the color of a person's shirt (*e.g.* red) can drastically switch under $G(s)$ (*e.g.* to blue) as long as $F(G(S))$ is able to reverse this process (see Fig. 8). This semantic shift corrupts the training data, since a synthetic image and its corresponding domain translated variant could look very different (*e.g.* the labels are not consistent). Semantic shift can occur because the *cycle-consistency* loss does not regulate the amount by which the domains can be shifted.

As mentioned in [52], one can adopt the technique from [39] and introduce an additional loss that forces the network to learn an identity mapping when samples from the target domain are provided as input to the generator, *i.e.* $\mathcal{L}_{id}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}[||G(x)-x||_1] + \mathbb{E}_{s\sim p_{data}(s)}[||F(s)-s||_1]$. Although, this loss helps to some degree, many subjects still exhibited drastic shifts in appearance.

Alternatively, we can integrate the loss from [37] which ensures the translated synthetic image is not too different from the original synthetic image *i.e.* $\mathcal{L}_{Ref}(G) = \mathbb{E}_{s\sim p_{data}(s)}[||G(s)-s||_1]$. We found this loss often leads to artifacts in the translated synthetic images, since the regularization does not distinguish between background/foreground. In practice, only the appearance of the person needs to be preserved. The background of synthetic image could be very different than what appears in the real images captured by the target camera.

To circumvent this issue, we apply a masking function which forces the network to focus on the foreground region

$$\mathcal{L}_{Mask}(G) = \mathbb{E}_{s\sim p_{data}(s)}\Big[\,\big\|\big(G(s)-s\big)*\mathbf{m}\big\|_1\Big], \tag{5}$$

where $\mathbf{m}$ is a mask that encourages the mapping to preserve the appearance only near to the center (see Fig. 5). Because re-identification datasets have well cropped images, the foreground region is typically in the middle of the bounding box, with the background around the periphery. Therefore, we pre-define a soft matte that resembles a 2D Gaussian kernel.

Our full objective loss is

$$\begin{aligned}\mathcal{L}_{our}(G,F,D_S,D_R) = {}& \mathcal{L}_{GAN}(G,D_R,S,R) + \mathcal{L}_{GAN}(F,D_S,R,S)\\ &+ \lambda_1\mathcal{L}_{cyc}(G,F) + \lambda_2\mathcal{L}_{id}(G,F) + \lambda_3\mathcal{L}_{Mask}(G),\end{aligned} \tag{6}$$

where $\lambda_1 = \lambda_2 = 10$ and $\lambda_3 = 5$ in our experiments (See Fig. 6).

Fine-Tuning. Given our re-identification network (see Sect. 4.1), we can fine-tune its feature extraction process to specialize for images generated from $G(s)$, which is our approximation of data coming from target domain (test camera). In practice, when we need to fine-tune our representation to a set of cameras, for every camera we identify its closest synthetic domain S_{k*} through our illumination inference, and then use it to learn a generator network that can transfer synthetic images to the given camera domain. The transferred synthetic images $G(s) : s \in S_{k*}$ are then used for fine-tuning the re-identification network, thus maximizing the performance of $\varPhi(G(s))$.

5 Experiments

We carried out experiments on 5 datasets: **VIPeR** [13], **iLIDS** [50], **CUHK01** [24], **PRID2011** [17] and **Market-1501** [48]. To learn a generic feature extrac-

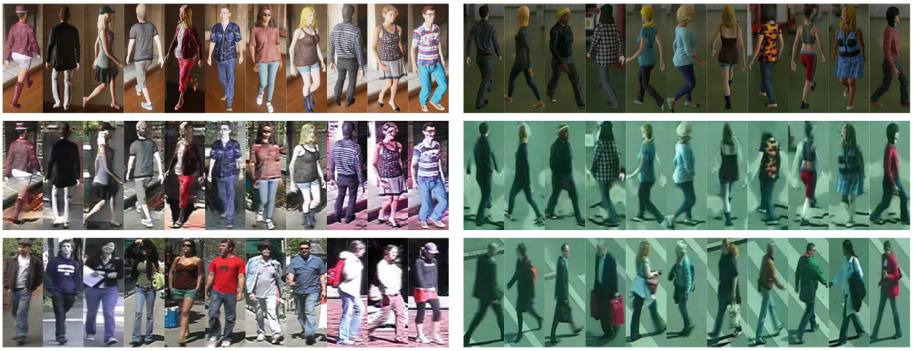

Fig. 6. Domain translation results for VIPeR (left) and PRID (right) datasets. From top to bottom: domain images $s \in S_{k*}$, translated images $G(s)$, target images $x \in R_{M+1}$.

tor we used two large scale re-identification datasets: **CUHK03** [25] and **DukeMTMC4ReID** [12,33], and our **SyRI** dataset. Re-identification performance is reported using rank-1 accuracy of the CMC curve [13].

Datasets: VIPeR contains 632 image pairs of pedestrians captured by two outdoor cameras. Large variations in lighting conditions, in background and in viewpoint are present. PRID2011 consists of person images recorded from two non-overlapping static surveillance cameras. Characteristic challenges of this dataset are extreme illumination conditions. There are two camera views containing 385 and 749 identities, respectively. Only 200 people appear in both cameras. i-LIDS consists of 476 images with 119 individuals. The images come from airport surveillance cameras. This dataset is challenging due to many occlusions. CUHK01 consists of $3,884$ images of 971 identities. There are two images per identity, per camera. The first camera captures the side view of pedestrians and the second camera captures the front or back view. Market-1501 contains 1501 identities, registered by at most 6 cameras. All the images were cropped by an automatic pedestrian detector, resulting in many inaccurate detections.

Evaluation Protocol: We generate probe/gallery images accordingly to the settings in [43]: VIPeR: 316/316; CUHK01: 486/486; i-LIDS: 60/60; and PRID2011: 100/649, where we follow a single shot setting [31]. For Market-1501 we employ the protocol from [44], where 750 test identities are used in a single query setting.

5.1 Generalization Properties

In this experiment, we train two feature extractors: one with only real images **R** containing CUHK03 and DukeMTMC4ReID images (in total 3279 identities); and the other one with both real and our synthetic images $\mathbf{R} + \mathbf{S}$ (our **SyRI** dataset provides additional 100 identities but under 140 illumination

Table 1. CMC rank-1 accuracy. The base model **R** is only trained on real images from auxiliary re-identification datasets. Adding synthetic images **S** improves the performance. Fine-tuning ($\mathbf{R} + \mathbf{S}^*$) to the training data of a specific dataset implies the maximum performance that could be expected with the correct synthetic data. Adapting the synthetic data to the target domain leads to significant gains, depending on the combination of semantic shift regularizations. Compared with state-of-the-art unsuperivsed techniques, our approach yields significantly higher accuracy on 4 of the 5 datasets. We achieve competitive performance to state-of-the-art on CUHK01

	Method	VIPeR	CUHK01	iLIDS	PRID	Market
Unsupervised	State-of-the-art	38.5 [42]	**57.3** [44]	49.3 [32]	34.8 [42]	58.2 [42]
	R	32.3	41.6	51.0	7.0	44.7
	$\mathbf{R} + \mathbf{S}$	36.4	49.5	54.8	15.0	54.3
	CycleGan	37.0	49.9	53.9	33.0	55.4
	CycleGan+$\mathcal{L}_{id}$	39.9	54.0	55.9	40.0	63.1
	CycleGan+$\mathcal{L}_{Ref}$	41.1	48.4	56.1	28.0	57.5
	Ours	**43.0**	54.9	**56.5**	**43.0**	**65.7**
	$\mathbf{R} + \mathbf{S}^*$	49.4	71.4	63.2	65.0	83.9

conditions, for a total of 3379 identities). For **S** we used 4 randomly sampled images per illumination condition per identity, which results in 56,000 images ($4 \times 140 \times 100$). Table 1 reports the performance comparison of these models on various target datasets. First, we evaluate the performance of the models directly on the target datasets without fine-tuning (fully unsupervised scenario, compare rows **R** and $\mathbf{R} + \mathbf{S}$, respectively). Adding our synthetic dataset significantly increases the re-identification performance. The row marked with * are the results after fine-tuning on the actual target datasets (*e.g.* in VIPeR column we fine-tune the model only on VIPeR dataset). It represents the maximum performance we expect to achieve if we could somehow hallucinate the perfect set of domain translated synthetic training images. These results indicate that the performance of supervised methods (using additional real data directly from the target domain) is still significantly better than unsupervised methods using domain adaptation. Interestingly, although adding our synthetic dataset doubled the performance on PRID2011, the lighting conditions in this dataset are so extreme that the gap to the supervised model is still significant. Similar findings have been reported in [2,42].

5.2 Illumination Inference

We carry out experiments to evaluate the importance of the illumination inference step. To do so, we compare the proposed *illumination estimator* to a random selection of the target illumination condition S_{k*}. After the illumination condition is selected, the proposed domain translation is applied. Table 2 illustrates the comparison on multiple dataset. We report minimum performance obtained by random procedure (MIN), the average across 10 experiments (Random),

Table 2. Impact of illumination inference. The selection of the right illumination condition for the domain translation improves the recognition performance

METHOD	VIPeR	CUHK01	iLIDS	PRID	Market
$\mathbf{R+S}$	36.4	49.5	54.8	15.0	54.3
MIN	35.2	50.4	55.1	29.0	58.1
RANDOM	38.9	51.2	56.1	36.0	60.9
Our	**43.0**	**54.9**	**56.5**	**43.0**	**65.7**

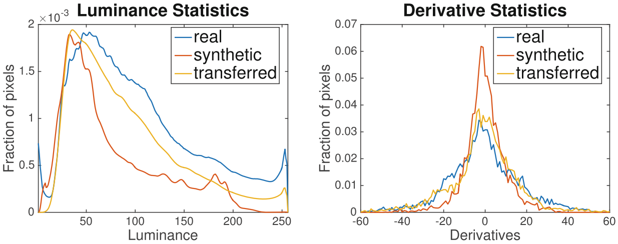

Fig. 7. Comparison of image statistics. Domain translation decreases the gap between synthetic and real image statistics.

and the average using our illumination inference. The results demonstrate that reasoning about illumination greatly improves the recognition performance. Our illumination condition estimator ensures that the source illumination is the closest to the target domain, thus facilitating the domain translation task.

5.3 Image Statistics of SyRI

The effect of domain translation is reflected in the underlying image statistics (see Fig. 7). The statistics of real and synthetic images are derived from a single camera from the VIPeR dataset and its corresponding camera in our SyRI dataset (selected by illumination inference). After passing through the generator function learned during domain translation ($G(s)$), the statistics of the translated images are much closer to the statistics of real images.

5.4 Domain Adaptation

Table 1 reports the performance of *CycleGAN* with different regularization terms. Domain translation without any regularization term between s and $G(s)$ can deteriorate performance (compare $\mathbf{R+S}$ and CYCLEGAN for iLIDS). We suspect this is due to the previously mentioned semantic shift (see Fig. 8). Adding identity mapping $\mathcal{L}_{id}$ makes significant improvement on both visual examples and re-identification performance. Replacing $\mathcal{L}_{id}$ with $\mathcal{L}_{Ref}$ can lower performance and tends to produce artifacts (notice artificial green regions in Fig. 8 for CUHK01). For CUHK01 and PRID datasets there are significant drops

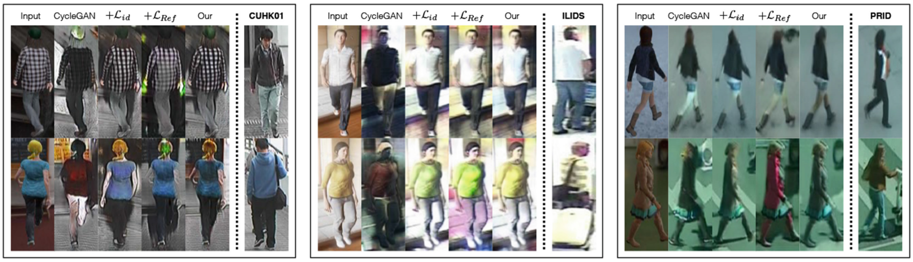

Fig. 8. Comparison of different regularization terms for translating synthetic images to a target domain. Representative image pairs for CUHK01, iLIDS and PRID datasets have been selected. Notice that CycleGAN without any regularization tends to have semantic shifts, *e.g.* for CUHK01 blue color of the t-shirt changed to red.

in the performance when using $\mathcal{L}_{Ref}$ regularization. Unlike [37], our images have distinct foreground/background regions. Background is not useful for re-identification, and it's influecen in the loss function should be minimial. Incorporating our mask makes significant improvements—especially for datasets where images are less tightly cropped, such as PRID. In this case, adding synthetic data improved performance from 7% to 15%. Our domain adaptation technique boosts the performance to 43.0% rank1-accuracy. We surpass the current state-of-the-art results by 8.2%.

5.5 Comparison with State-of-the-Art Methods

We divide the state-of-the-art approaches into unsupervised and supervised techniques as well as methods that employ hand-crafted features (including graph-learning GL [20] and transfer learning TL [20]) and embeddings learned with Convolutional Neural Newtworks (CNN) (including source identity knowledge transfer learning CAMEL [44] and attribute knowledge transfer TJ-AIDL [42]). Table 3 illustrates that: (1) our three-step domain adaptation technique outperforms the state-of-the-art unsupervised techniques—on 4 of the 5 datasets, we outperform the state-of-the-art results by large margins: 5.1%, 7.2%, 8.2% and 7.5% on VIPeR, iLIDS, PRID and Market, respectively; on CUHK01 we achieve competitive performance to CAMEL [44] (2.4% performance gap), but CAMEL performs significantly worse than our approach on VIPeR and Market. Compared with other augmentation techniques (*e.g.* SPGAN [6]), our illumination inference step ensures that the source illumination, chosen from a large number of options in our SyRI dataset, is closest to the target domain. (2) When compared to unsupervised hand-crafted based approaches, the performance margins for rank-1 are even larger: 11.5%, 13.9%, 7.2% and 18% on VIPeR, CUHK01, iLIDS and PRID, respectively. (3) Our approach is also very competitive with the best supervised techniques—regardless of the dataset. This confirms the effectiveness of the proposed solution, which does not require any human supervision and thus scales to large camera networks.

Table 3. Comparison with state-of-the-art unsupervised and supervised techniques. The best scores for unsupervised methods are shown in **bold**. The best scores of supervised methods are highlighted in red

		Method	VIPeR	CUHK01	iLIDS	PRID	Market
Unsupervised	Hand-craft	GL [20]	33.5	41.0	–	25.0	–
		DLLAP [21]	29.6	28.4	–	21.4	–
		TSR [35]	27.7	23.3	–	–	–
		TL [32]	31.5	27.1	49.3	24.2	–
	CNN	SSDAL [38]	37.9	–	–	20.1	39.4
		CAMEL [44]	30.9	**57.3**	–	–	54.5
		SPGAN [6]	–	–	–	–	57.7
		TJ-AIDL [42]	38.5	–	–	34.8	58.2
		Ours	**43.0**	54.9	**56.5**	**43.0**	**65.7**
Supervised	Hand-craft	LOMO+XQDA[27]	40.0	63.2	–	26.7	–
		Ensembles[31]	45.9	53.4	50.3	17.9	–
		Null Space[45]	42.2	64.9	–	29.8	55.4
		Gaussian+XQDA [29]	49.7	57.8	–	–	66.5
	CNN	Triplet Loss[4]	47.8	53.7	60.4	22.0	–
		FT-JSTL+DGD[43]	38.6	66.6	64.6	64.0	73.2
		SpindleNeT[46]	**53.8**	**79.9**	**66.3**	**67.0**	**76.9**

6 Conclusion

Re-identification datasets contain many identities, but rarely have a substantial number of different lighting conditions. In practice, this lack of diversity limits the generalization performance of learned re-identification models on new unseen data. Typically, the networks must be fine-tuned in a supervised manner using data collected for each target camera pair, which is infeasible at scale. To solve this issue, we propose a new synthetic dataset of virtual people rendered in indoor and outdoor environments. Given example unlabelled images from a test camera, we develop an illumination condition estimator to select the most appropriate subset of our synthesized images to use for fine-tuning a pre-trained re-identification model. Our approach is ideal for large scale deployments, since no labelled data needs to be collected for each target domain.

We employ a deep network to modify the subset of synthesized images (selected by the illumination estimator) so that they more closely resemble images from the test domain (see Fig. 6). To accomplish this, we use the recently introduced cycle-consistent adversarial architecture and integrate an additional regularization term to ensure the learned domain shift (between synthetic and real images) does not result in generating unrealistic training examples (*e.g.* drastic changes in color). Because re-identification images have distinct foreground/background regions, we also incorporate a soft matte to help the network focus on ensuring the foreground region is correctly translated to the target domain. Extensive experiments on multiple datasets (see Table 3) show that our approach outperforms other unsupervised techniques, often by a large margin.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
2. Bak, S., Carr, P.: One-shot metric learning for person re-identification. In: CVPR, June 2017
3. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR, July 2017
4. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: CVPR, June 2016
5. Debevec, P.: Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In: Proceedings of ACM SIGGRAPH, pp. 189–198 (1998)
6. Deng, W., Zheng, L., Kang, G., Yang, Y., Ye, Q., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
7. Dibra, E., Maye, J., Diamanti, O., Siegwart, R., Beardsley, P.: Extending the performance of human classifiers using a viewpoint specific approach. In: WACV (2015)
8. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV (2013)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015)
10. Gao, S., Tsang, I.W.H., Chia, L.T., Zhao, P.: Local features are not lonely–laplacian sparse coding for image classification. In: CVPR (2010)
11. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
12. Gou, M., Karanam, S., Liu, W., Camps, O., Radke, R.J.: DukeMTMC4ReID: a large-scale multi-camera person re-identification dataset. In: CVPRW (2017)
13. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS (2007)
14. Hattori, H., Boddeti, Y.V.N., Kitani, K.M., Kanade, T.: Learning scene-specific pedestrian detectors without real data. In: CVPR (2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, June 2016
16. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arxiv (2017)
17. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21227-7_9
18. Hu, J., Lu, J., Tan, Y.P.: Deep transfer metric learning. In: CVPR (2015)
19. Huang, S., Ramanan, D.: Expecting the unexpected: training detectors for unusual pedestrians with adversarial imposters. In: CVPR (2017)
20. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Person re-identification by unsupervised ℓ_1 graph learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 178–195. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_11

21. Kodirov, E., Xiang, T., Gong, S.: Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In: BMVC (2015)
22. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR, July 2017
23. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR, June 2018
24. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_3
25. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
26. Li, Z., Chang, S., Liang, F., Huang, T., Cao, L., Smith, J.: Learning locally-adaptive decision functions for person verification. In: CVPR (2013)
27. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
28. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. In: CVPR, June 2014
29. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: CVPR, June 2016
30. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: SceneNet RGB-D: can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In: ICCV, October 2017
31. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR (2015)
32. Peng, P., et al.: Unsupervised cross-dataset transfer learning for person re-identification. In: CVPR, June 2016
33. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
34. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
35. Shi, Z., Hospedales, T.M., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: CVPR (2015)
36. Shotton, J., et al.: Efficient human pose estimation from single depth images. TPAMI **35**(12), 2821–2840 (2013)
37. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
38. Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q.: Deep attributes driven multi-camera person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 475–491. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_30
39. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: arXiv preprint (2016)
40. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
41. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: ICCV (2015)

42. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
43. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR (2016)
44. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV (2017)
45. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR (2016)
46. Zhao, H., et al.: Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: CVPR (2017)
47. Zheng, L., et al.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52
48. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)
49. Zheng, W.S., Gong, S., Xiang, T.: Towards open-world person re-identification by one-shot group-based verification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 591–606 (2016)
50. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
51. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR (2011)
52. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)