Instance-Guided Context Rendering for Cross-Domain Person Re-Identification

Yanbei Chen
Queen Mary University of London
yanbei.chen@qmul.ac.uk

Xiatian Zhu
Vision Semantics Ltd.
eddy.zhuxt@gmail.com

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Abstract

Existing person re-identification (re-id) methods mostly assume the availability of large-scale identity labels for model learning in any target domain deployment. This greatly limits their scalability in practice. To tackle this limitation, we propose a novel Instance-Guided Context Rendering scheme, which transfers the source person identities into diverse target domain contexts to enable supervised re-id model learning in the unlabelled target domain. Unlike previous image synthesis methods that transform the source person images into limited fixed target styles, our approach produces more visually plausible, and diverse synthetic training data. Specifically, we formulate a dual conditional generative adversarial network that augments each source person image with rich contextual variations. To explicitly achieve diverse rendering effects, we leverage abundant unlabelled target instances as contextual guidance for image generation. Extensive experiments on Market-1501, DukeMTMC-reID and CUHK03 benchmarks show that the re-id performance can be significantly improved when using our synthetic data in cross-domain re-id model learning.

1. Introduction

Person re-identification (re-id) is a task of re-identifying a query person-of-interest, across non-overlapping cameras distributed over wide surveillance spaces [16]. Since the surge of deep representation learning, great boosts of re-id performance have been witnessed in an idealistic closed-world supervised learning testbed [63, 58, 54, 64, 20, 6, 30, 47, 5]: The rank-1 matching rate has reached 93.3% [5] on the Market1501 benchmark [63], as compared to 44.4% in 2015. However, this success relies heavily on an *unrealistic* assumption that the training and test data have to be drawn from the same camera network, i.e. the same domain. When deploying such re-id models to new domains, their performances often degrade significantly, mainly due to the inevitable domain gaps between datasets collected from different surveillance camera networks. This weakness greatly restricts the generalisability of these *domain-specific* learning methods in real-world deployment, when manually labelling new identity population becomes prohibitively expensive at large scale [57, 11, 55, 28, 7, 34]. It is therefore essential to automate the domain-adaptive learnability with more advanced and robust *domain-generic* learning models.

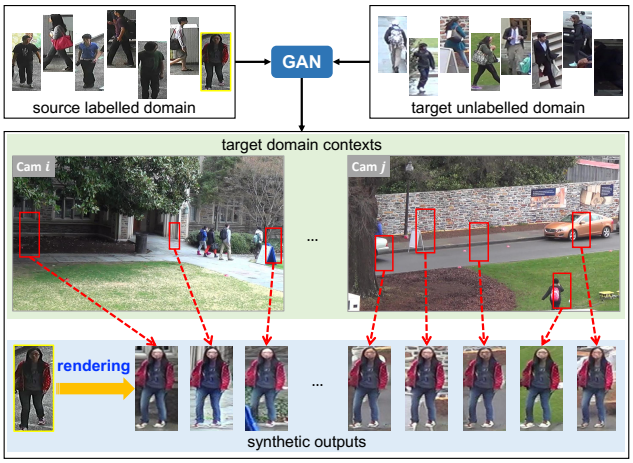


Figure 1: Motivation illustration. In open surveillance spaces, the *contextual variations* can be quite diverse, due to *wide-of-the-field imagery* and *varying times of the day*. Our approach learns to hallucinate the same persons in such surveillance contexts, as if they were captured from different places and times in the target domain.

The aforementioned problem, known as cross-domain person re-id, is gaining increasing attention [40, 56, 37, 57, 11, 55, 1, 65, 33]. It raises a more challenging *open-set* unsupervised domain adaptation problem [4, 44], which requires to bridge the domain gap between two *disjoint* identity class spaces. Recent methods typically mitigate this gap by attribute-identity distribution alignment at the feature level [55, 33], or style transfer at the image level [57, 11]. However, they all neglect to exploit the *rich contextual variations* as a potential domain bridge. In this work, we aim to utilise the contextual information for more effective re-id model learning. This is motivated by our observation of complex environmental dynamics commonly existed in open public scenes (see Fig. 1) – domain contexts are indeed quite diverse in surveillance spaces, given that the viewing

conditions vary dramatically both *within* and *across* camera views, subjected to camera characteristics, wide-field-of-view imagery, and varying times of the day. Our key idea is to render the source persons into diverse domain contexts, such that a large-scale *context augmented synthetic dataset* can be generated to train a re-id model in a supervised manner without labelling any target domain data.

Specifically, we propose a novel Instance-Guided Context Rendering scheme, which augments the same source identity population with rich contextual variations reflected in the target domain. Our approach is unique in several perspectives. *First*, it effectively exploits *abundant unlabelled target instances* as guidance to render the source persons into different target domain contexts. This essentially captures the image-level domain drift in a more comprehensive way. *Second*, rather than optimising two-way mappings heavily with cycle consistency, we learn a simple *one-way mapping* through informative supervision signals. *Third*, compared to previous GAN-based re-id methods [57, 11], our proposed *dual conditional* formulation naturally avoids *mode collapse* [2] to limited styles, and enables more diverse outputs. It transfers the same person into more realistic, finer-grained, and richer viewing conditions. The contextually more diverse synthetic imagery are ultimately utilised for re-id model learning to enhance visual invariance towards contextual variations in the target domain.

In summary, our **contribution** is two-fold:

- We propose a novel Instance-Guided Context Rendering scheme. To our best knowledge, it is *the first attempt in re-id* to tackle the image-level domain drift by injecting *rich contextual information* into the image generation process. It effectively augments the same source person images with diverse target domain contexts to construct a large-scale synthetic training set for re-id model learning in the unlabelled target domain.
- We design a dual conditional generative adversarial network. It effectively exploits abundant unlabelled target instances as contextual guidance to produce more plausible data with richer *cross-* and *intra-domain contextual variations*. We conduct extensive experiments to validate our model design rationale, and show that our approach not only achieves competitive re-id performance on several re-id benchmarks in the cross-domain setting, but also generates photo-realistic person images with high fidelity and diversity.

2. Related Work

Unsupervised Cross-Domain Person Re-Identification aims to transfer the identity discriminative knowledge from a labelled source domain to an unlabelled target domain. The state-of-the-art methods [57, 11, 55, 65, 33, 1, 31] can be categorised into three learning paradigms: (1) *feature-level distribution alignment*; (2) *image-level style transfer*; and (3) *hybrid image-level and feature-level learning*. The first paradigm [55, 33] generally seeks a common feature space for source-target distribution alignment with discriminative learning constraints. The second paradigm [57, 11, 1, 31] reduces the domain gap by using GAN frameworks to transfer source images into target domain styles in a holistic manner. The last paradigm [65] unifies the complementary benefits of synthetic images by GAN and feature discriminative constraints in CNN. Our work falls into the second paradigm. In particular, we identify that the common weakness of existing GAN-based re-id methods lies in the insufficient data diversity – either *one* or a *pre-fixed* number of domain styles are captured in the final outputs. This is mainly caused by the *mode collapse* issue in GAN – very limited styles are plausibly captured in generated outputs. To rectify this weakness, we design a new GAN framework to augment data with more diverse contexts. Our synthetic images reflect richer contextual variations in the target domain, and naturally serve as more informative training data to improve the domain generalisability of a re-id model.

Unsupervised Domain Adaptation (UDA) techniques [51, 49, 36, 46, 14, 50, 3, 45, 48, 53, 18, 22, 59] aim to tackle the domain drift for avoiding exhaustive manual labelling of target data. Existing UDA methods rely on either *feature-level adaptation* [51, 36, 46, 14, 50, 59] or *image-level adaptation* [3, 45] to mitigate the cross-domain distribution discrepancy. The former focuses on learning domain-invariant feature representation, which is generally achieved by adversarial training [14, 50, 53], or aligning the feature statistics, such as sample means [51, 59] and covariances [46]. The latter seeks to stylise the source images to look visually as the target domain images using generative models [3, 45, 48]. Rooted in similar spirit, our approach also learns to transform the image styles, with a particular focus on enriching the diversity of synthetic images to facilitate more effective domain adaptation in re-id.

Image-to-Image Translation (I2I) aims to transform images from original styles to new styles [24, 66, 25, 35, 60, 9, 38, 8, 23, 27]. The first unified I2I framework Pix2Pix [24] adopts conditional GAN to learn a one-way mapping by optimising GAN loss and reconstruction loss formed on pairwise labelled data. CycleGAN [66] utilises the cycle consistency constraints to avoid pairwise supervision by learning two-way cross-domain mappings. To further enable multi-domain mappings, StarGAN [9] introduces discrete domain labels as conditional variables to capture multiple modes. Recently, MUNIT [23], DRIT [27] achieve more diverse image translation by conditioning the generation on random latent codes. Driven by the same goal of diversifying generated outputs, we design a *dual conditional* formulation to augment richer contextual variations in person images for boosting cross-domain re-id model learning.

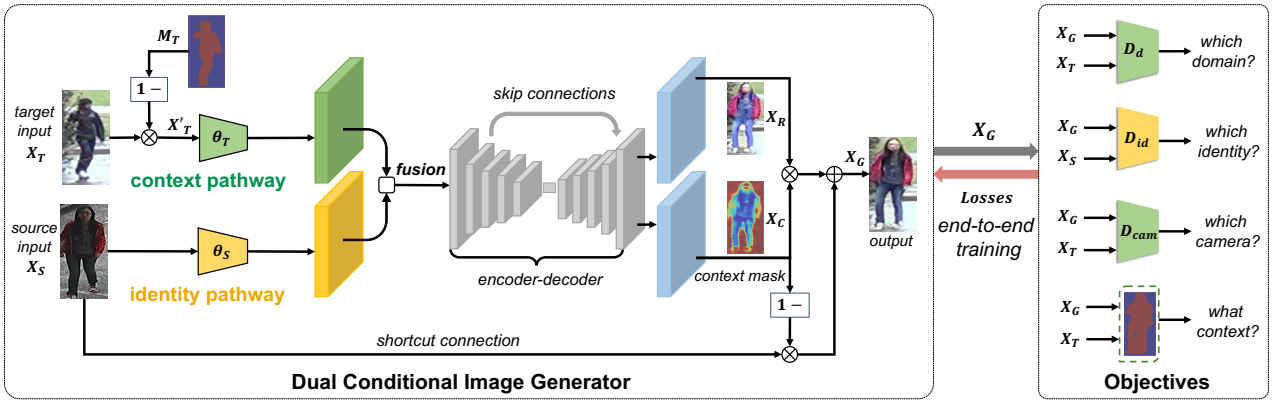


Figure 2: **Model overview.** We tackle the domain drift at the image level by learning to render the source person image $\boldsymbol{X}_S$ into diverse domain contexts explicitly guided by arbitrary target instances $\boldsymbol{X}_T$ sampled from the target domain (Sec. 3.1, Sec. 3.2).

Figure 3: **Deployment overview.** In deployment, the generator produces abundant synthetic images $\boldsymbol{X}_G$ for CNN training (Sec. 3.3).

3. Instance-Guided Context Rendering

Problem Definition. We consider the problem of unsupervised domain adaptation in person re-id, which aims to adapt a re-id model learned from a labelled source dataset to an unlabelled target dataset. Our objective is to learn a generative mapping G that reduces the domain discrepancy by rendering the same source person images into a diverse range of target domain contexts. As the final synthetic images are augmented with rich target contexts, a CNN model can be simply fine-tuned upon these data to enhance its generalisability in the unlabelled target domain.

Approach Overview. Fig. 2 illustrates our Instance-Guided Context Rendering scheme. Its main body is a dual conditional Generative Adversarial Network that takes in a pair of input images from two domains for image generation (Sec. 3.1), and learns with informative supervision signals to render the source persons guided by different target instances (Sec. 3.2). We name our **C**ontext **R**endering Network as **CR-GAN** for short. In deployment (Fig. 3), abundant data augmented with diverse context is exploited for re-id model learning in the synthetic target domain (Sec. 3.3).

3.1. Dual Conditional Image Generator

Dual Conditional Mapping. CR-GAN contains a dual conditional image generator that learns a one-way mapping to render the source images into desired target contexts by conditioning on two inputs: a source input $\boldsymbol{X}_S$ and a target input instance $\boldsymbol{X}_T$ to guide the context rendering effect. Formally, this dual conditional mapping is expressed as:

$$\boldsymbol{X}_G = G(\boldsymbol{X}_S, \boldsymbol{X}_T) \tag{1}$$

In essence, this dual conditional formulation is designed to fuse the information flows from two domains, such that the same person in source input $\boldsymbol{X}_S$ can be rendered into the target context explicitly guided by the target instance $\boldsymbol{X}_T$. Overall, the whole mapping is built upon dual-path encoding and decoding, with a U-Net [43] like encoder-decoder network in between, as detailed below.

Dual-Path Encoding. To enable instance-guided context rendering, we introduce an essential condition $\boldsymbol{X}_T$ to exploit abundant target instances as contextual guidance in image generation. Concretely, we design a dual-path encoding structure to parameterise information flows from two domain separately (Fig. 2) – **(1)** An *identity pathway* θ_S to encode source input $\boldsymbol{X}_S$; and **(2)** A *context pathway* θ_T to encode target input $\boldsymbol{X}_T$. Given that our aim is to exploit contextual information from target domain, we mask the target input $\boldsymbol{X}_T$ to retain mainly the background clutter. Specifically, we adopt the off-the-shelf human parsing model LIP-JPPNet [15] to obtain a binary person mask, and apply spatial masking on $\boldsymbol{X}_T$ to filter out the target person:

$$\boldsymbol{X}'_T = \boldsymbol{X}_T \circ (\boldsymbol{1} - \boldsymbol{M}_T) \tag{2}$$

where $\circ$ is the Hadamard product; $\boldsymbol{M}_T$ is the person mask of input $\boldsymbol{X}_T$; $\boldsymbol{X}'_T$ contains mainly the background clutter.

Through dual-path encoding, the information flows from two domains are further fused by depth-wise concatenation: $[\theta_S(\boldsymbol{X}_S), \theta_T(\boldsymbol{X}'_T)]$, followed with an encoder-decoder network to selectively blend the visual information from two inputs. We construct the encoder-decoder network as a cascade of up-sampling, down-sampling residual blocks, along with skip connections that enforce the generator network to selectively preserve low-level visual structures from both conditional inputs. In particular, the foreground person in $\boldsymbol{X}_S$, the background clutter in $\boldsymbol{X}_T$ should both be picked by the generator as informative cues for image generation.

Image Generation. To render the context in a region-selective manner, i.e. keeping the source person whilst augmenting background clutters, we employ a context mask to softly specify the region of contextual changes. Concretely, the generator outputs two parts: **(1)** A *residual map* $\boldsymbol{X}_R$ to model cross-domain discrepancy; and **(2)** A *context mask* $\boldsymbol{X}_C$ to modulate per-pixel intensity of context change, both of which are connected by a shortcut connection to reuse the source person in input $\boldsymbol{X}_S$. Such generic masking mechanisms are also adopted in recent literature, such as face animation [41], motion manipulation [62]; while we particularly utilise the context mask to automatically learn the region selection of context rendering. The final generated output $\boldsymbol{X}_G$ is the sum of source input $\boldsymbol{X}_S$ and residual map $\boldsymbol{X}_R$ spatially weighted by the context mask $\boldsymbol{X}_C$:

$$\boldsymbol{X}_G = \boldsymbol{X}_R \circ \boldsymbol{X}_C + \boldsymbol{X}_S \circ (\mathbf{1} - \boldsymbol{X}_C) \tag{3}$$

The generator is trained end-to-end to generate $\boldsymbol{X}_G$, which retains the person identity as $\boldsymbol{X}_S$ in the new context of $\boldsymbol{X}_T$.

3.2. Learning Objectives

The key idea of CR-GAN is to inject context information into image generation. This is motivated that contextual variations exist at multi-granularity – they not only differ across domains, but also vary dramatically within and across camera views. To learn such variations, we impose *four* distinct losses for model optimisation, which work synergistically to learn (a) *cross-domain*, (b) *cross-camera*, and (c) *inner-camera context variations*, whilst (d) retaining the *source identity*, as illustrated in Fig. 4 and detailed below.

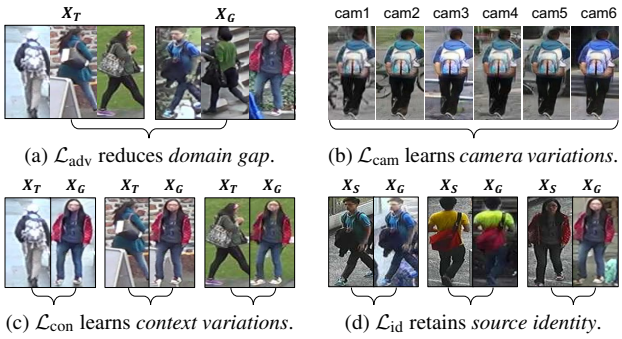
Adversarial Loss. To mitigate the *cross-domain* contextual gap, the generator G is trained against a domain discriminator D_d in an adversarial minimax manner [17]:

$$\mathcal{L}_{\text{adv}}=\min_{G}\max_{D_d} \log D_d(\boldsymbol{X}_T)+\log(1-D_d(G(\boldsymbol{X}_S,\boldsymbol{X}_T))) \tag{4}$$

where $\mathcal{L}_{\text{adv}}$ aligns the generated data distribution with the target data distribution globally to reduce the domain gap.

Camera Loss. To capture the *cross-camera* context variations induced by camera characteristics – e.g. *colour tones* – a camera loss is imposed to constrain the camera styles:

$$\mathcal{L}_{\text{cam}}=-\log(p(y_c|\boldsymbol{X}_G)) \tag{5}$$



(a) $\mathcal{L}_{\text{adv}}$ reduces *domain gap*. (b) $\mathcal{L}_{\text{cam}}$ learns *camera variations*.

(c) $\mathcal{L}_{\text{con}}$ learns *context variations*. (d) $\mathcal{L}_{\text{id}}$ retains *source identity*.

Figure 4: Schematic illustration of learning objectives.

where y_c is camera label of $\boldsymbol{X}_T$. $\mathcal{L}_{\text{cam}}$ is derived by a camera discriminator D_{cam} trained to classify camera labels.

Context Loss. Besides capturing context variations across domains and cameras, the generator should also learn the *inner-camera* context variations with content details. Accordingly, we adopt masked reconstruction errors to constrain the foreground, background same as input $\boldsymbol{X}_S, \boldsymbol{X}_T$:

$$\mathcal{L}_{\text{con}}=||(\boldsymbol{X}_G-\boldsymbol{X}_S)\circ\boldsymbol{M}_F||_2+||(\boldsymbol{X}_G-\boldsymbol{X}_T)\circ\boldsymbol{M}_B||_2 \tag{6}$$

where $\boldsymbol{M}_F, \boldsymbol{M}_B$ are the foreground, background person masks of $\boldsymbol{X}_S$ extracted by human parsing model. $\mathcal{L}_{\text{con}}$ particularly encourages to retain the source person, whilst augmenting more *diverse background clutters* explicitly guided by arbitrary target instance $\boldsymbol{X}_T$ from the target domain.

Identity Loss. As the source person identity in input $\boldsymbol{X}_S$ should be preserved in output $\boldsymbol{X}_G$, we impose an identity classification error to constrain the person identity in $\boldsymbol{X}_G$:

$$\mathcal{L}_{\text{id}}=-\log(p(y_j|\boldsymbol{X}_G)) \tag{7}$$

where y_j is the identity label of $\boldsymbol{X}_S$; $\mathcal{L}_{\text{id}}$ is derived by an identity discriminator D_{id} – a standard re-id CNN model trained to predict the source person identities.

Overall Objective. CR-GAN is trained with the joint optimisation of four losses (Eq. (4),(5),(6),(7)) for their complementary benefits in constraining the image generation:

$$\mathcal{L}_{GAN}=\lambda_{\text{adv}}\mathcal{L}_{\text{adv}}+\lambda_{\text{id}}\mathcal{L}_{\text{id}}+\lambda_{\text{cam}}\mathcal{L}_{\text{cam}}+\lambda_{\text{con}}\mathcal{L}_{\text{con}} \tag{8}$$

where $\lambda_{\text{adv}}, \lambda_{\text{id}}, \lambda_{\text{cam}}, \lambda_{\text{con}}$ are hyper-parameters to control the relative importance of each loss. We set $\lambda_{\text{id}}=\lambda_{\text{cam}}=1$, $\lambda_{\text{adv}}=2$, $\lambda_{\text{con}}=5$ to keep the losses in similar value range.

3.3. Model Training and Deployment

CR-GAN is optimised similar to standard GAN models, as summarised in Alg. 1. For deployment, D_{id} – a standard backbone ResNet50 [19] – is fine-tuned upon abundant synthetic data generated by CR-GAN (Fig. 3). All synthetic data is randomly produced on-the-fly by feeding arbitrary image pairs to CR-GAN, therefore eschewing the need of storing an extremely large-scale synthetic dataset. After fine-tuning, the backbone network D_{id} is deployed to extract feature for re-id matching in the target domain.

Algorithm 1 Algorithmic Overview.

I. Initialisation: Pre-train D_{id}, D_{cam} with labels.
II. Train the image generator G:
Input: Source dataset $\mathcal{D}_S$, target dataset $\mathcal{D}_T$.
Output: An image generator G.
for $t = 1$ **to** *max_gan_iter* **do**
  Feedforward mini-batch of input pairs $(\boldsymbol{X}_S, \boldsymbol{X}_T)$ to G.
  Update D_d (Eq. (4)) and update G for k times (Eq. (8)).
end for
III. Fine-tune D_{id} on synthetic data:
for $t = 1$ **to** *max_cnn_iter* **do**
  Random context rendering: $\boldsymbol{X}_G = G(\boldsymbol{X}_S, \boldsymbol{X}_T)$.
  Update D_{id} on $\boldsymbol{X}_G$ using identity label of $\boldsymbol{X}_S$.
end for

3.4. Discussion

Overall, our CR-GAN has several merits that can benefit cross-domain re-id model learning: **(1)** Instead of controlling the rendering effects with a fixed set of category labels [9], e.g. camera labels, we leverage *abundant unlabelled instances* $\boldsymbol{X}_T$ from target domain as contextual guidance to inject contextual variations. This naturally avoids *mode collapse* to limited fixed styles, and synthesises more diverse target domain contexts for learning a domain-generic re-id model. **(2)** Rather than changing the domain contexts holistically [23], our rendering effects are region-selective. In particular, the *background clutter* is modified significantly with *structural change*; while the *foreground person* is inpainted slightly with *colour change* to capture the domain drift. Such rendering effects effectively retain the source identity, whilst augmenting much richer contexts for re-id model learning in the synthetic target domain. **(3)** By fusing two inputs through dual-path encoding at the lower layers, the generator network is enforced to learn the selective preservation of low-level visual structures from both inputs, therefore enhancing the modelling capacity to produce synthetic training data in higher fidelity and diversity.

4. Experiments

4.1. Experimental settings

Implementation Details. To train CR-GAN, we use the Adam solver [26] with a mini-batch size of 32. The learning rate is set to 0.0002 in the first half of training and linearly decayed to 0 in the second half. To build up the image generator, Instance Normalisation (IN) [52] is used in the U-Net decoder. IN is neither applied in two separate encoding pathways nor the U-Net encoder, which allows to retain the stylistic information before decoding. The two pathways for dual condition are parameterised as separate convolutional layers. To improve the training stability of GAN, we add one additional Gaussian noise layer as the input layer in the domain discriminator. We employ LSGAN [39] as the GAN formulation and adopt the domain discriminator same as PatchGAN [24] to discriminate at the scale of patches. To stabilise the training, the image generator is updated twice every iteration in the second half of training. We use the standard ImageNet [10] pre-trained ResNet50 as the identity discriminator D_{id}. The camera discriminator D_{cam} is an extremely lightweight CNN classifier with 5 layers. The image generator G, domain discriminator D_d are iteratively updated as shown in Alg. 1. After training, the ResNet50 is used as the backbone network to extract feature for re-id evaluation. More details on network architectures and training procedures are given in the Supplementary Material.


(a) Market1501. (b) DukeMTMCreID. (c) CUHK03.
Figure 5: Example images from three re-id benchmarks.

Evaluation Metrics. We adopt several metrics to comprehensively evaluate our model in two aspects. **(1)** To evaluate the re-id matching performance, we adopt the standard *Cumulative Match Characteristic* (**CMC**) and *mean Average Precision* (**mAP**) as evaluation metrics. We report results on *single-query* based on the ranking order of cross-camera pairwise matching distances computed using features extracted from the re-id CNN model. **(2)** To measure the visual quality of synthesis, we adopt the following two evaluation metrics: **(i)** *LPIPS Distance* (**LPIPS**) [61] measures the *image translation diversity*, which is correlated with human perceptual similarity. We use the default ImageNet pre-trained AlexNet to extract feature in evaluation. **(ii)** *Fréchet Inception Distance* (**FID**) [21] measures the *image fidelity* by quantifying the distribution discrepancy between generated data and real data. We use the default ImageNet pre-trained Inception to extract feature in evaluation.

Datasets. We adopt three standard re-id benchmarks for evaluation (Fig. 5). **(1) Market1501** [63] contains 1,501 identities captured by 6 different cameras. The training set includes 751 identities and 12,936 images. The test set includes 750 identities, with 3,368 images in the probe set and 19,732 images in the gallery set. **(2) DukeMTMCreID** [42, 64] contains 1,404 identities captured by 8 different cameras. The training set includes 702 identities and 16,522 images. The testing set includes 702 identities, with 2,228 images in the probe set and 17,661 images in the gallery set. **(3) CUHK03** [29] contains 1,467 identities and 14,097 images in total. We use the auto-detected version.

4.2. Ablative Model Evaluation

To validate our model design rationale, we first conduct ablation study on two different domain pairs: Market1501 $\rightarrow$ DukeMTMCreID, DukeMTMCreID $\rightarrow$ Market1501.

Figure 6: Qualitative visual evaluation. Given source image $\boldsymbol{X}_S$, (a) baseline (w/o dual condition) *collapses* to uniform context, due to lack of *contextual guidance*; while (b) CR-GAN augments the same person with diverse contexts explicitly guided by target instances $\boldsymbol{X}_T$.

S → T	Market→Duke		Duke→Market	
Metrics	LPIPS	FID	LPIPS	FID
Source-Target data	0.458	0.330	0.458	0.330
w/o dual condition	0.196	0.065	0.210	0.137
CR-GAN	**0.281**	**0.058**	**0.269**	**0.096**

Table 1: Quantitative visual evaluation on image quality. **LPIPS**: *image perceptual similarity*, higher is better. **FID**: *distribution discrepancy*, lower is better. LPIPS / FID in "Source-Target data" represents the upper bound. Best results are in **bold**.

S → T	Market→Duke		Duke→Market	
Metrics (%)	R1	mAP	R1	mAP
Direct Transfer	36.9	20.5	47.5	20.0
w/o dual cond	43.3	24.8	55.5	27.0
CR-GAN	**52.2**	**30.0**	**59.6**	**29.6**
w/o dual cond+LMP	48.7	27.6	59.2	28.5
CR-GAN+LMP	**56.0**	**33.3**	**64.5**	**33.2**

Table 2: Ablation study of dual condition in re-id. "Direct Transfer": CNN trained with only labelled source data; "w/o dual cond": without dual condition; LMP: a pooling strategy [11] to reduce noisy signals induced by fake synthetic images at test time.

S → T	Market→Duke		Duke→Market	
Metrics (%)	R1	mAP	R1	mAP
w/o identity loss	31.9	15.4	32.8	11.8
w/o camera loss	48.8	28.6	53.6	26.0
w/o context loss	48.5	28.8	57.4	28.7
CR-GAN	**52.2**	**30.0**	**59.6**	**29.6**

Table 3: Ablation study on individual effect of each loss in re-id.

Effect of Dual Condition. Introducing abundant target instances as contextual guidance is the *key factor* that enables an Instance-Guided Context Rendering process. To validate this factor, we compare our dual conditional mapping (CR-GAN) with an ablative baseline that takes in merely the source input $\boldsymbol{X}_S$ (w/o dual condition). Fig. 6 shows that: **(1)** Although the baseline transforms the context, all the generated images *collapse* to the same context; **(2)** CR-GAN, on the contrary, acts as a much stronger data generator to augment the same person with a more diverse range of domain contexts. This is in line with our visual quantitative results in Table 1, where CR-GAN obtains much higher LPIPS, i.e. more diverse outputs, compared to the baseline. This shows compellingly the benefit of our *dual conditional* formulation to exploit abundant target instances as contextual guidance in the image generation.

To evaluate the benefit of context rendering effects in re-id, we compare CR-GAN with the ablative baseline. Table 2 shows that **(1)** Introducing our dual conditional formulation significantly boosts the re-id performance, with improved margins of 8.9% (52.2-43.3) / 4.1% (59.6-55.5) in R1 on DukeMTMCreID / Market1501. **(2)** The improvement remains in the use of LMP, with improved margins of 7.3% (56.0-48.7) / 5.3% (64.5-59.2) in R1. This indicates that re-id model learning with more contextual variations is indeed helpful to boost the cross-domain model robustness.

Effect of Different Losses. In addition to the standard adversarial loss, CR-GAN is trained with *three* different losses. To validate the necessity of using these losses in re-id, we conduct ablative comparison by eliminating individual loss from the overall objective. Table 3 shows that: **(1)** Removing any of the loss leads to undesired performance drop; **(2)** All losses work synergistically, with their joint optimisation to achieve the best performance. **(3)** These results are in line with our loss design rationale: All losses serve to exploit the complementary information in model optimisation (Fig. 4), thus giving their desired performance gains to yield better synthetic data for re-id model learning.

4.3. Analysis on GAN-based Methods

To isolate and analyse the pure effect of image-level domain adaptation in re-id, we compare our model with GAN-based methods for ablative analysis in this section.

Qualitative Visual Analysis. To understand how context information is brought to benefit the re-id model learning,

Figure 7: Qualitative visual evaluation. Given source image $\boldsymbol{X}_S$, (a) SPGAN [11] transforms the image into merely *one uniform style*; while (b) our CR-GAN renders the source persons into varying contexts: different *background clutters*, *colour tones* and *lighting conditions*.

S → T	Market→Duke		Duke→Market	
Metrics	LPIPS	FID	LPIPS	FID
Source-Target data	0.458	0.330	0.458	0.330
SPGAN [11]	0.099	0.171	0.099	0.115
CR-GAN	**0.281**	**0.058**	**0.269**	**0.096**

Table 4: Quantitative visual evaluation on image quality. **LPIPS**: *image perceptual similarity*, higher is better. **FID**: *distribution discrepancy*, lower is better. Best results are in **bold**.

S → T	Market→Duke		Duke→Market	
Metrics (%)	R1	mAP	R1	mAP
PTGAN [57]	27.4	-	38.6	66.1
SPGAN [11]	41.1	22.3	51.5	22.8
M2M-GAN [31]	49.6	26.1	57.5	26.8
CR-GAN	**52.2**	**30.0**	**59.6**	**29.6**
SPGAN+LMP [11]	46.4	26.2	57.7	26.7
M2M-GAN+LMP [31]	54.4	31.6	63.1	30.9
CR-GAN+LMP	**56.0**	**33.3**	**64.5**	**33.2**

Table 5: Evaluation on GAN-based methods in the cross-domain re-id settings. Best results in each group are in **bold**. Overall $1^{st}/2^{nd}$ best in red/blue.

we first visually compare the synthetic images produced by our CR-GAN with SPGAN [11]: a *representative* re-id method based upon *CycleGAN*. As Fig. 7 shows, compared to merely one plausible output given by SPGAN, CR-GAN can produce more diverse outputs. This informs that CR-GAN indeed serves as a much stronger *synthetic data generator* to augment much more contextual variations and thus produces a synthetic training set of much larger-scale.

Quantitative Visual Analysis. To evaluate the visual quality quantitatively, we further compare CR-GAN with SPGAN based on the synthetic data released by the authors. Table 4 indicates that: **(1)** Both CR-GAN and SPGAN have lower and better FID compared to the FID between the source and target data. This informs that after style adaptation, the cross-domain distribution discrepancy is mitigated with both methods. **(2)** Compared to SPGAN, CR-GAN has much lower FID and higher LPIPS. This indicates CR-GAN can generate images of better fidelity and higher diversity.

Analysis on Re-id Matching. To further justify how our synthetic contextual variations benefit cross-domain re-id learning, we compare CR-GAN with three state-of-the-art GAN-based re-id methods: PTGAN [57], SPGAN [11], M2M-GAN [31] on two domain pairs. All these models are trained on the *same source datasets* under the *same learning paradigm*: a GAN is first trained to synthesise images, a CNN is then fine-tuned upon the synthetic data for domain adaptation. Table 5 shows that CR-GAN achieves the best cross-domain re-id performance. It is worth pointing out that previous methods generally collapse to fixed style(s): one homogenous domain style (PTGAN, SPGAN), or a predefined set of camera styles (M2M-GAN). In contrast, CR-GAN augments much richer contextual variations that ultimately benefit domain adaptation in re-id.

4.4. Comparison with the State-of-the-art

Competitors. We compare our CR-GAN with 12 state-of-the-art methods. To ensure a ***like-to-like fair comparison***, we compare these methods by categorising them into four groups: **(a)** *shallow methods using hand-crafted features*: LOMO, BoW, UMDL; **(b)** *image-level learning methods*: PTGAN, SPGAN, M2M-GAN, which use GANs for style transfer; **(c)** *feature-level learning methods*: PUL, TJ-AIDL, MMFA, BUC, TAUDL, which use additional discriminative constraints in CNN; **(d)** *hybrid learning methods*: HHL, which combine the benefits of group **(b)** and **(c)**.

It is worth noting that the learning paradigms in group **(b)**, **(c)** are essentially *orthogonal*: learning is performed either in *image space* or *feature space*. Therefore, these two paradigms should be complementary when unified in a hybrid formulation. To testify the generalisability of CR-GAN in a hybrid formulation, we add an additional comparison by unifying CR-GAN / SPGAN with the best performer TAUDL in group **(c)**. We first train the CNN with synthetic data generated by CR-GAN / SPGAN, then apply TAUDL with the pre-trained CNN for unsupervised learning in the target domain. Such hybrid formulations are denoted as CR-GAN+TAUDL / SPGAN+TAUDL, respectively.

Evaluation on Market1501 / DukeMTMCreID. Table 6

Types	Source → Target	Market1501 → DukeMTMCreID				DukeMTMCreID → Market1501			
	Metrics (%)	R1	R5	R10	mAP	R1	R5	R10	mAP
Shallow	LOMO [32]	12.3	21.3	26.6	4.8	27.2	41.6	49.1	8.0
	BoW [63]	17.1	28.8	34.9	8.3	35.8	52.4	60.3	14.8
	UMDL [40]	18.5	31.4	37.6	7.3	34.5	52.6	59.6	12.4
Image	PTGAN [57]	27.4	-	50.7	-	38.6	-	66.1	-
	SPGAN+LMP [11]	46.4	62.3	68.0	26.2	57.7	75.8	82.4	26.7
	M2M-GAN+LMP [31]	54.4	-	-	31.6	63.1	-	-	30.9
	CR-GAN+LMP	**56.0**	**70.5**	**74.6**	**33.3**	**64.5**	**79.8**	**85.0**	**33.2**
Feature	PUL* [13]	30.0	43.4	48.5	16.4	45.5	60.7	66.7	20.5
	TJ-AIDL† [55]	44.3	59.6	65.0	23.0	58.2	74.8	81.1	26.5
	MMFA† [33]	45.3	59.8	66.3	24.7	56.7	75.0	81.8	27.4
	BUC* [34]	47.4	62.6	68.4	27.5	**66.2**	79.6	84.5	38.3
	TAUDL* [28]	**61.7**	-	-	**43.5**	63.7	-	-	**41.2**
Hybrid	HHL [65]	46.9	61.0	66.7	27.2	62.2	78.8	84.0	31.4
	SPGAN+TAUDL	66.1	80.0	83.2	47.2	66.5	81.8	86.6	38.5
	CR-GAN+TAUDL	**68.9**	**80.2**	**84.7**	**48.6**	**77.7**	**89.7**	**92.7**	**54.0**

Table 6: Evaluation on Market1501, DukeMTMCreID in comparison to the state-of-the-art unsupervised cross-domain re-id methods. *: Not use auxiliary source training data. †: Use auxiliary source attribute labels for training. "-": no reported results. Best results in each group are in **bold**. Overall $1^{st}/2^{nd}$ best in red/blue. Note that HHL uses StarGAN [9] to generate synthetic training images.

Types	Source → Target	CUHK03 → Market1501				CUHK03 → DukeMTMCreID			
	Metrics (%)	R1	R5	R10	mAP	R1	R5	R10	mAP
Image	PTGAN [57]	31.5	-	60.2	-	17.6	-	38.5	-
	SPGAN [11]	42.3	-	-	19.0	-	-	-	-
	CR-GAN	**58.5**	**75.8**	**81.9**	**30.4**	**46.5**	**61.6**	**67.0**	**26.9**
Feature	TAUDL* [28]	**63.7**	-	-	**41.2**	**61.7**	-	-	**43.5**
Hybrid	HHL [65]	56.8	74.7	81.4	29.8	42.7	57.5	64.2	23.4
	CR-GAN+TAUDL	**78.3**	**89.4**	**93.0**	**56.0**	**67.7**	**79.4**	**83.4**	**47.7**

Table 7: Evaluation on CUHK03 to Market1501 / DukeMTMCreID adaption compared to state-of-the-art unsupervised cross-domain re-id methods. *: Not use source data. "-": no reported results. Best results in each group are in **bold**. Overall $1^{st}/2^{nd}$ best in red/blue.

shows comparative results on two domain pairs. It can be observed that **(1)** CR-GAN performs best in the *image-level learning* paradigm; **(2)** When deploying CR-GAN in a hybrid formulation (CR-GAN+TAUDL), we earn the best re-id performance due to the complementary benefits of two learning paradigms. In particular, CR-GAN+TAUDL boosts the performance over TAUDL with margins of 7.2% (68.9-61.7) / 14.0% (77.7-63.7) in R1 on DukeMTMCreID / Market1501. These results not only indicate the benefit of unifying ***GAN-based*** *image-level learning* and ***CNN-based*** *feature-level learning* into unsupervised cross-domain re-id, but more importantly justify our rationale of augmenting richer contextual variations to enable learning a more effective re-id model in the applied domain.

Evaluation on CUHK03 to Market1501 / DukeMTMCreID. Table 7 shows comparative results on model adaptation from CUHK03, where there exists larger domain gaps between the source and target domains (Fig. 5). It can be seen that **(1)** CR-GAN clearly outperforms the best image-level competitor SPGAN with large margins; **(2)** When deploying in a hybrid formulation, CR-GAN+TAUDL outperforms the best hybrid competitor HHL with large margins of 21.5% (78.3-56.8), 25.0% (67.7-42.7) in R1 on Market1501 / DukeMTMCreID respectively. These collectively suggest the significant advantages of exploiting the synthetic data by CR-GAN in cross-domain re-id model learning.

5. Conclusion

We presented a novel Instance-Guided Context Rendering scheme for cross-domain re-id model learning. Through a carefully-designed dual conditional mapping, abundant target instances are exploited as contextual guidance for image generation. We conducted extensive ablative analysis to validate our model design rationale, and show the best performance over existing GAN-based re-id methods. Our like-to-like comparison with the state-of-the-art methods demonstrates the great advantage of our model when flexibly deploying in a hybrid systematic formulation. Overall, CR-GAN serves as a generic generator to augment abundant domain contexts for re-id model learning in practice.

Acknowledgements This work is supported by Vision Semantics Limited, the China Scholarship Council, the Alan Turing Institute, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[2] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Pixelnn: Example-based image synthesis. In *International Conference on Learning Representation*, 2018. 2

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[4] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763, 2017. 1

[5] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *European Conference on Computer Vision*, 2018. 1

[6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *In IEEE International Conference on Computer Vision Workshops*, 2017. 1

[7] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. In *British Machine Vision Conference*, 2018. 1

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Combogan: Unrestricted scalability for image domain translation. In *Workshop of International Conference on Learning Representation*, 2018. 2

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 5, 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 6, 7, 8

[12] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision*, 2017. 13

[13] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 8

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016. 2

[15] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

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

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 4

[18] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *IEEE International Conference on Computer Vision*, 2017. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, 2018. 2

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, 2018. 2, 5

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 2, 5

[25] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, 2017. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2014. 5

[27] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2

[28] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *European Conference on Computer Vision*, 2018. 1, 8

[29] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 5

[30] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[31] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Junyong Zhu. M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. In *AAAI Conference on Artificial Intelligence*, 2019. 2, 7, 8

[32] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 8

[33] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *British Machine Vision Conference*, 2018. 1, 2, 8

[34] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, 2019. 1, 8

[35] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, 2017. 2

[36] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015. 2

[37] Andy Jinhua Ma, Jiawei Li, Pong C Yuen, and Ping Li. Cross-domain person reidentification using domain adaptation ranking svms. *IEEE Transactions on Image Processing*, 2015. 1

[38] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[39] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 5

[40] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 8

[41] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision*, 2018. 4

[42] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 5

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 3

[44] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *European Conference on Computer Vision*, 2018. 1

[45] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[46] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016. 2

[47] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *European Conference on Computer Vision*, 2018. 1

[48] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representation*, 2017. 2

[49] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, 2015. 2

[50] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[51] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[52] Dmitry Ulyanov, Andrea Vedaldi, and Victor S Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[53] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[54] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[55] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 8

[56] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 1

[57] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 7, 8

[58] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[59] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2018. 2

[60] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision*, 2017. 2

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[62] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *European Conference on Computer Vision*, 2018. 4

[63] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015. 1, 5, 8

[64] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision*, 2017. 1, 5

[65] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *European Conference on Computer Vision*, 2018. 1, 2, 8

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 2