Learning Disentangled Representation for Robust Person Re-identification (ID 2853)

We thank all the reviewers for their valuable comments. We will clarify their concerns in the paper.

FD-GAN (R1). FD-GAN and IS-GAN are similar in that both use a GAN-based distillation technique for a robust reID. Differently, FD-GAN extracts identity-related and pose-unrelated features, but with extra pose labels. Distilling other factors except for human pose is not feasible. In contrast, IS-GAN disentangles identity-related and -unrelated features through identity shuffling, factorizing other factors irrelevant to person reID, such as pose, scale, background clutter, and occlusion, without supervisory signals for them. Accordingly, the identity-related feature for IS-GAN is much more robust to such factors of variations than the identity-related and pose-unrelated feature for FD-GAN, and this gives a superior performance on the Market-1501 and DukeMTMC-reID datasets. Note that CUHK03 was excluded, as FD-GAN used a different training/test split.

MGN (R1). MGN uses the same backbone network as IS-GAN to extract initial part-level features. As it is trained with a hard-triplet loss, the part-level features are highly discriminative, but they are not robust to *e.g.*, pose, scale, background clutter, and occlusion. MGN thus shows the reID performance comparable with IS-GAN on Market-1501, where discriminative attributes of identities can be captured well. For example, person images with the same identity are almost identical in the dataset. MGN, however, shows a limited performance on the CUHK03 and DukeMTMC-reID datasets, where the same person is captured with different poses, view points, background, and occlusion.

DG-Net (R2). DG-Net (CVPR 2019) was not published at the time of our submission. It thus should not be our consideration, but we'd like to clarify here the difference from DG-Net. Although appearance/structure features in DG-Net seem to be analogous to identity-related/-unrelated ones in IS-GAN, they are completely different. DG-Net computes appearance/structure features by AdaIN (ICCV 2017), widely used in image stylization, and thus they are more like style/content features. Figure 9 in Appendix of the DG-Net paper visualizes generated person images when structure features (analogous to identity-unrelated features of IS-GAN) are changed only. We can see that DG-Net even changes the entire attributes (*e.g.*, gender) except the color information, suggesting that structure features also contain id-related cues. Note that IS-GAN outperforms DG-Net for all benchmarks by a large margin (*e.g.*, rank-1/mAP on DukeMTMC-reID: 90.0/78.1 (IS-GAN) and 86.6/74.8 (DG-Net)).

Table 1: Quantitative comparison for a different network.

		Market-1501	
	$\mathcal{L}_{\mathrm{PS}}$	R-1	mAP
PCB	✗	91.0	74.2
PCB + IS-GAN	✗	<u>92.4</u>	<u>77.2</u>
PCB + IS-GAN	✓	**92.7**	**77.5**

IS-GAN with a different backbone (R2). To evaluate the generalization ability, we tried to use PCB as our backbone to extract CNN features, and added IS-GAN on top of the features. We modified the network architecture such that each part-level feature has the size of $1 \times 1 \times 256$ for an efficient computation, and set this as our baseline. Note that the original PCB also gives six part-level features, but with the size of $1 \times 1 \times 2,048$ for each feature. Table 1 shows that our method improves the baseline consistently, suggesting it can be applied to other methods.

The number of body parts (R2) We show in Table 2 the effect of the part-level shuffling loss on the different number of body parts. We can see that 1) the part-level shuffling loss generalizes well across the different number of body parts, and 2) IS-GAN shows better performance as more body parts are used.

Table 2: Ablation studies on the number of body parts.

		Market-1501	
	$\mathcal{L}_{\mathrm{PS}}$	R-1	mAP
part-2	✗	84.1	61.3
	✓	88.8	68.7
part-3	✗	86.9	65.7
	✓	91.2	74.2
part-1,2	✗	91.9	77.8
	✓	92.1	78.2
part-1,3	✗	93.4	81.1
	✓	93.7	81.3

More results for disentangled features. (R3). Figure 1 shows an example of generated images using a part-level identity shuffling technique. This corresponds to Fig. 5 in the main paper, but with *different identities*, demonstrating once again that IS-GAN successfully disentangles identity-related and -unrelated features in a part-level. For example, we can see, in the upper left picture, that IS-GAN changes colors of T-shirts between persons, while preserving the poses and background. On the contrary, colors of T-shirts are maintained, while the poses and background are changed in the upper right picture.

Hyperparameter (R1). We empirically found that training with a large value of λ_{U} is unstable. We thus set λ_{U} to 0.001 in the second stage, and increased to 0.01 in the third stage to regularize the disentanglement. We used a grid search to set other parameters with $\lambda_{\mathrm{R}} \in \{5, 10, 20\}$, $\lambda_{\mathrm{PS}} \in \{5, 10, 20\}$, and $\lambda_{\mathrm{C}} \in \{1, 2\}$ on the Market-1501 dataset. We randomly split IDs in the training dataset into 651/100 and used corresponding images as training/validation sets. Following [27, 35], we fixed λ_{S} and λ_{D} to 10 and 1, respectively. We fixed all parameters and trained our model on the CUHK03 and DukeMTMC-reID datasets.

Discriminators (R3). The domain and class discriminators share five blocks consisting of `conv-instnorm-lrelu`, and each has an independent head (L217-L218). For the domain discriminator, we added two more blocks, resulting in a features map of size 12×4. We then used this as an input to PatchGAN. For the class discriminator, we added one more block followed by a fully connected layer. At the time of the publication, we will make our source code and models open to the public.

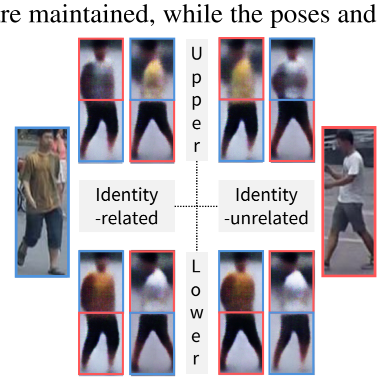

Figure 1: Visualization of disentangled features for person images with different identities.