Joint Discriminative and Generative Learning for Person Re-identification

Zhedong Zheng, Xiaodong Yang, Zhiding Yu,
Liang Zheng, Yi Yang, Jan Kautz

박태우

Introduction

- Person Re-ID의 주요 문제는 다른 카메라 간의 intra-class variations.

- Intra-class variations의 영향을 줄이기 위해 part-based matching, part alignment 와 같은 방법을 이용.

- 다른 방법으로는 GAN을 이용한 data augmentation으로 학습을 진행.

- 기존의 연구에서는 image generation을 Re-ID learning과 분리를 시킴.

- 본 논문에서는 위의 두 learning을 한꺼번에 학습시키는 unified framework를 제안.

Introduction

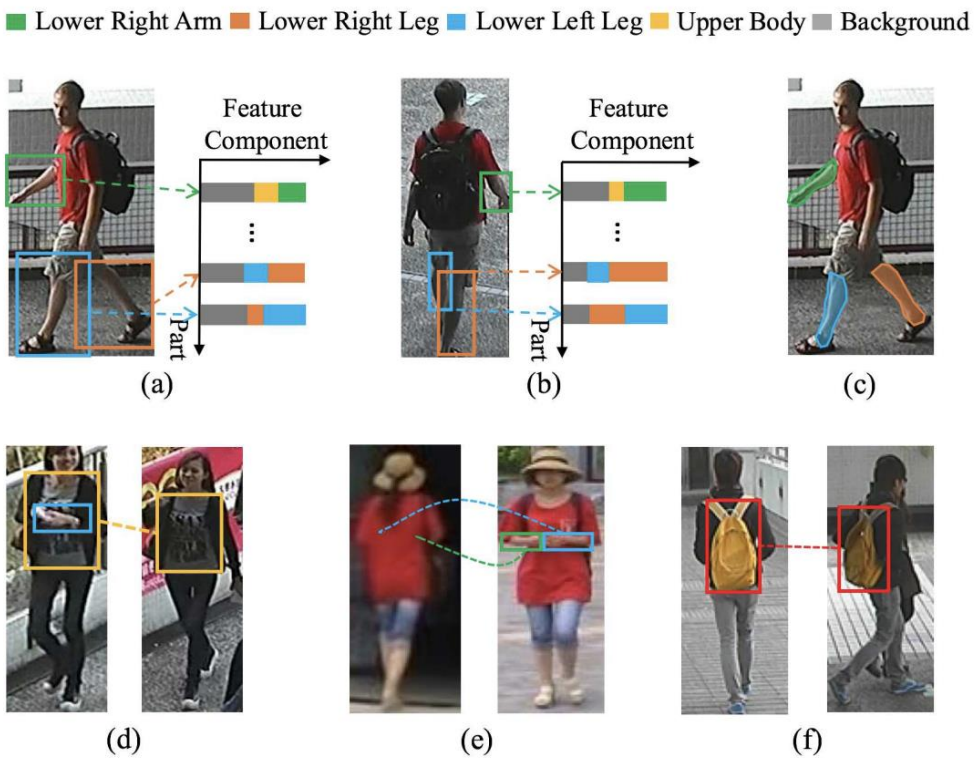
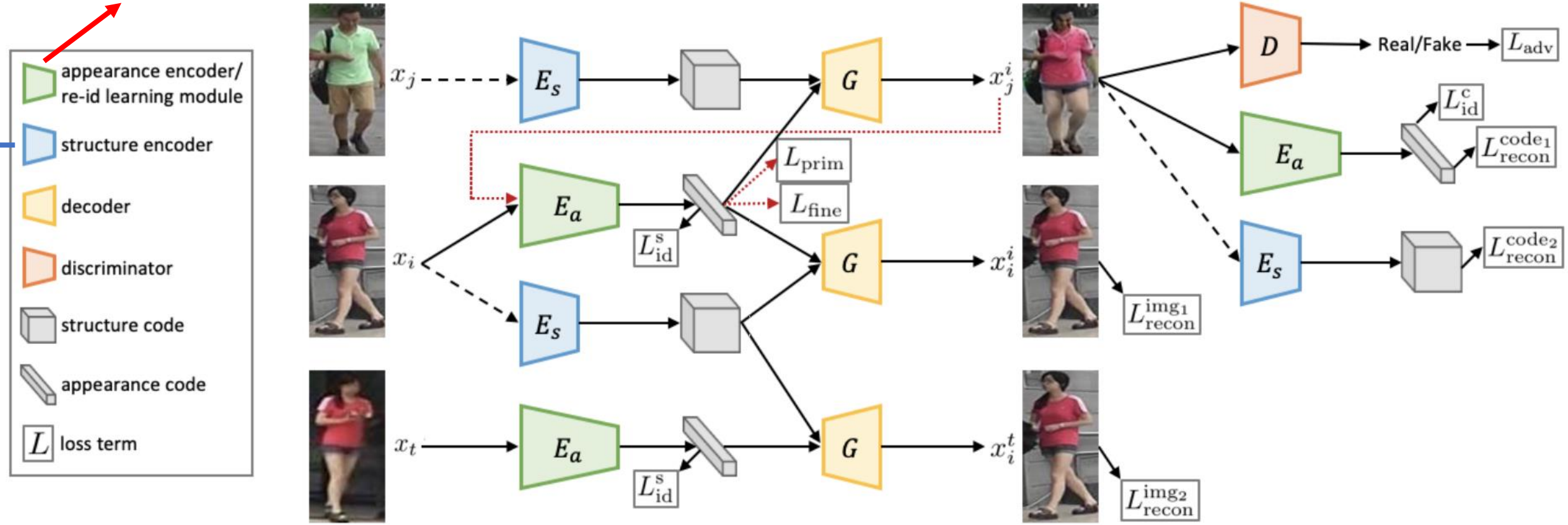


Figure 1. Part alignment challenges in person ReID. (a-c): Describing body parts by bounding boxes may introduce many irrelevant regions from background and other parts. Matching between features extracted from loose boxes in (a) and tight boxes in (b) would deteriorate the matching accuracy. A finer part region representation in (c) would help alleviate this problem. (d-f): The importance of different parts should be adaptively adjusted. Upper body is occluded by forearms in (d), the two forearms are all occluded in (e). Features from occluded part should be eliminated during matching, while salient visual cues like yellow backpack in (f) need to be emphasized.

Architecture

DG-Net

ResNet50 이용, 즉 확률분포의 값이 output임.



Sturucture 정보를 얻기 때문에 Appearance encoder보다 spatial함. 하지만 decoder에서 보통 spatial한 정보에서 feature를 얻는 경향이 있기 때문에 gray scale을 사용.

Generative module

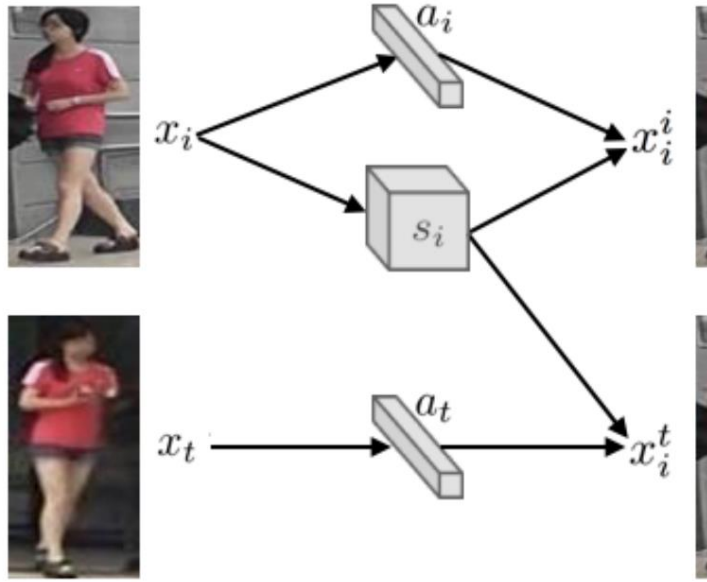
1. Self-identity generation

Reduce intra-class feature variations

Original image 1

Generated image 1

Original image 1

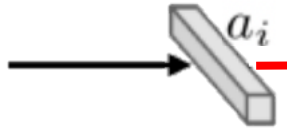


$$L_{\text{recon}}^{\text{img}_1} = \mathbb{E}[\|x_i - G(a_i, s_i)\|_1]$$

$$L_{\text{recon}}^{\text{img}_2} = \mathbb{E}[\|x_i - G(a_t, s_i)\|_1]$$

Original image 2 $(y_i = y_t)$ Generated image 2 Original image 1

Image



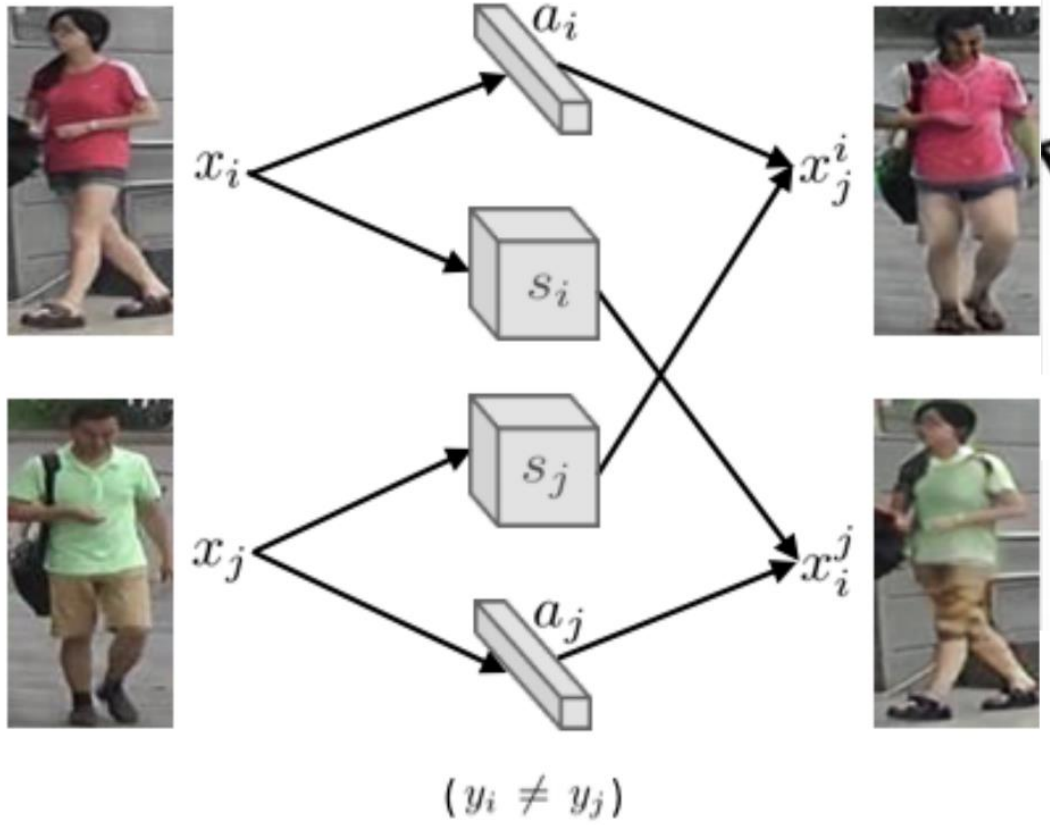
Appearance code

$$L_{\text{id}}^{\text{s}} = \mathbb{E}[-\log(p(y_i|x_i))]$$

Identification loss

Generative module

2. Cross-identity gerneration



$$L_{\text{adv}} = \mathbb{E}[\log D(x_i) + \log(1 - D(G(a_i, s_j))]$$

D → Real/Fake

$$L_{\text{id}}^{\text{c}} = \mathbb{E}[-\log(p(y_i|x_j^i))]$$

E_a → a_i

$$L_{\text{recon}}^{\text{code}_1} = \mathbb{E}[\|a_i - E_a(G(a_i, s_j))\|_1]$$

E_s → s_j

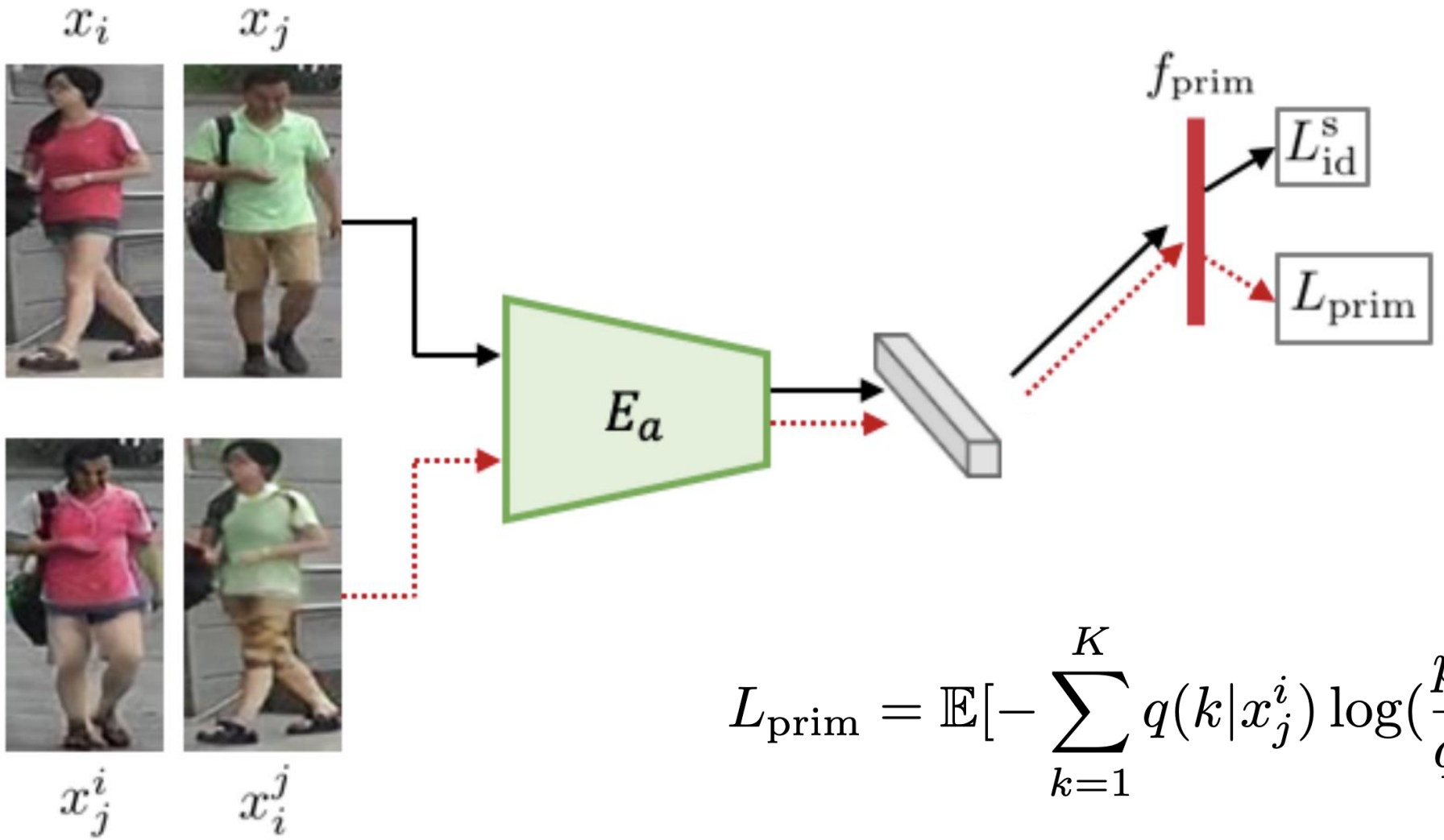
$$L_{\text{recon}}^{\text{code}_2} = \mathbb{E}[\|s_j - E_s(G(a_i, s_j))\|_1]$$

Discriminative module

1. Primary feature learning

서로 다른 이미지 2개의 합성 이미지를 training data로 사용

Teacher – Student 구조의 supervision learning 이용.



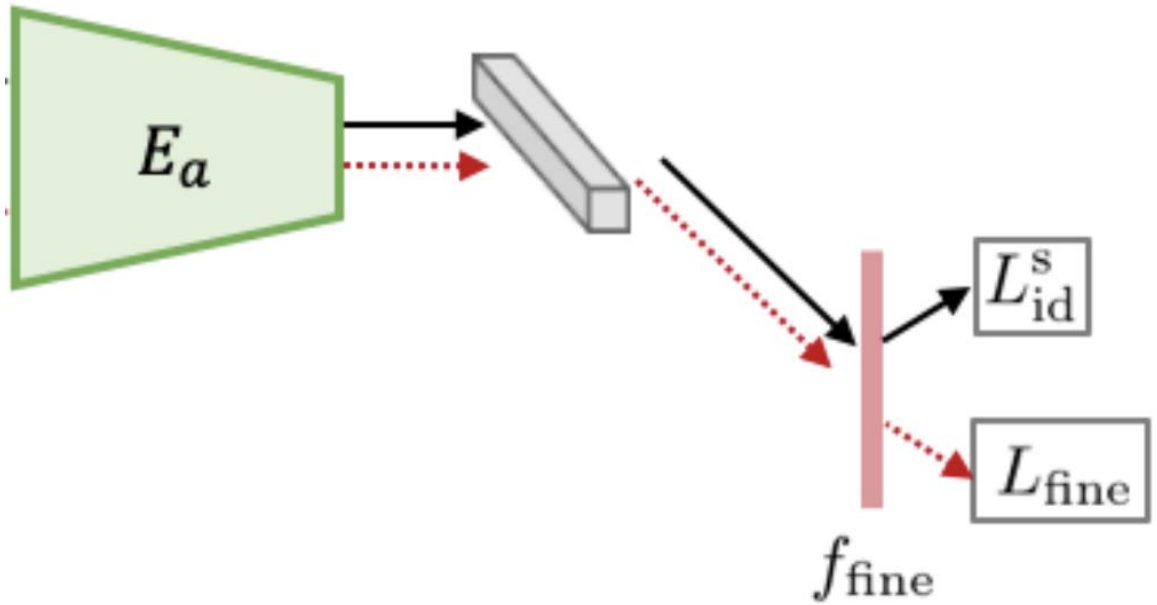
$$L_{\text{prim}} = \mathbb{E}[-\sum_{k=1}^{K} q(k|x_j^i) \log(\frac{p(k|x_j^i)}{q(k|x_j^i)})]$$

Discriminative module

2. Fine-grained feature mining

1개 이미지를 appearance만 계속 바꿔 학습을 진행.

$$L_{\text{fine}} = \mathbb{E}[-\log(p(y_j|x_j^i))]$$



Optimization

$$L_{\text{total}}(E_a, E_s, G, D) = \lambda_{\text{img}} L_{\text{recon}}^{\text{img}} + L_{\text{recon}}^{\text{code}} + L_{\text{id}}^{\text{s}} + \lambda_{\text{id}} L_{\text{id}}^{\text{c}} + L_{\text{adv}} + \lambda_{\text{prim}} L_{\text{prim}} + \lambda_{\text{fine}} L_{\text{fine}},$$

각 요소에 가중치를 곱해 최종 합산을 함.

Result

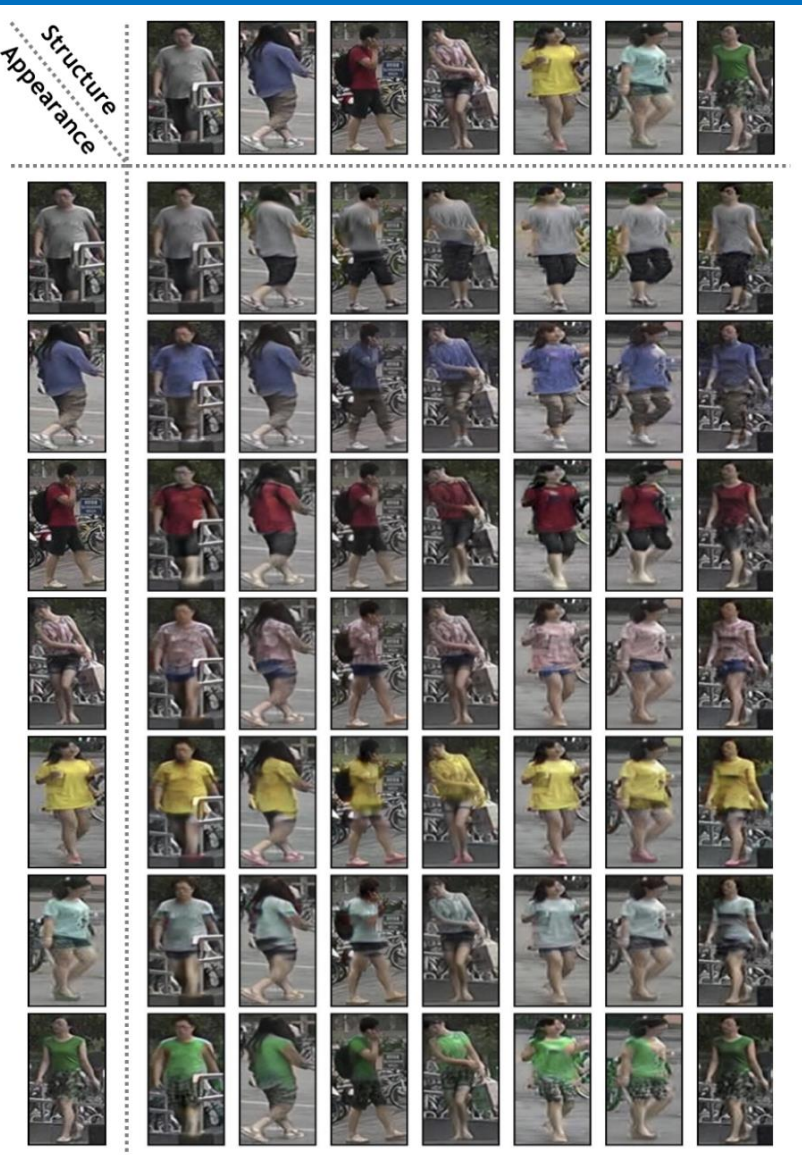


Figure 1: Examples of generated images on Market-1501 by switching appearance or structure codes. Each row and column corresponds to different appearance and structure.

Result

Figure 3: Comparison of the generated and real images on Market-1501 across the different methods including LSGAN [29], PG^2-GAN [28], FD-GAN [10], PN-GAN [31], and our approach. This figure is best viewed when zoom in. Please attention to both foreground and background of the images.

Result

Figure 6: Examples of our generated images by swapping appearance or structure codes on the three datasets. All images are sampled from the test sets.

Result

Methods	Realism (FID)	Diversity (SSIM)
Real	7.22	0.350
LSGAN [29]	136.26	-
PG^2-GAN [28]	151.16	-
PN-GAN [31]	54.23	0.335
FD-GAN [10]	257.00	0.247
Ours	**18.24**	**0.360**

Table 2: Comparison of FID (lower is better) and SSIM (higher is better) to evaluate realism and diversity of the real and generated images on Market-1501.

Result

Figure 7: Comparison of success and failure cases in our image generation. In the failure case, the logo on t-shirt of the original image is missed in the synthetic image.

Result

Methods	Market-1501		DukeMTMC-reID	
	Rank@1	mAP	Rank@1	mAP
Verif-Identif [55]	79.5	59.9	68.9	49.3
DCF [22]	80.3	57.5	-	-
SSM [2]	82.2	68.8	-	-
SVDNet [38]	82.3	62.1	76.7	56.8
PAN [57]	82.8	63.4	71.6	51.5
GLAD [47]	89.9	73.9	-	-
HA-CNN [24]	91.2	75.7	80.5	63.8
MLFN [4]	90.0	74.3	81.0	62.8
Part-aligned [37]	91.7	79.6	84.4	69.3
PCB [39]	93.8	81.6	83.3	69.2
Mancs [43]	93.1	82.3	84.9	71.8
DeformGAN [34]	80.6	61.3	-	-
LSRO [56]	84.0	66.1	67.7	47.1
Multi-pseudo [17]	85.8	67.5	76.8	58.6
PT [27]	87.7	68.9	78.5	56.9
PN-GAN [31]	89.4	72.6	73.6	53.2
FD-GAN [10]	90.5	77.7	80.0	64.5
Ours	**94.8**	**86.0**	**86.6**	**74.8**

Table 4: Comparison with the state-of-the-art methods on the Market-1501 and DukeMTMC-reID datasets. Group 1: the methods not using generated data. Group 2: the methods using separately generated images.

Reference

http://openaccess.thecvf.com/content_cvpr_2018/papers/Xu_Attention-Aware_Compositional_Network_CVPR_2018_paper.pdf

https://bismex.github.io/2019/06/10/DGNET.html