Attribute-aware Pedestrian Image Editing [⋆]

Xiaoyi Yin[1,2], Xinqian Gu[1,2], Hong Chang[1,2], Bingpeng Ma[2], and Xilin Chen[1,2]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
{xiaoyi.yin, xinqian.gu}@vipl.ict.ac.cn, bpma@ucas.ac.cn,
{changhong,xlchen}@ict.ac.cn

Abstract. Pedestrian image generation is a very challenging task. Existing generation methods have drawbacks including body distortion, inadequate visual details and large vague areas. In this paper, we propose Attribute-aware Pedestrian Image Editing (APIE) to address these problems based on given visual attributes. Our model denominated as APIE-Net, has three mechanisms including an attribute-aware segmentation network, a multi-scale discriminator and a latent-variable discriminator. Experiments on Market-1501 and DukeMTMC-reID datasets show that APIE-Net can generate satisfying pedestrian images with given attributes. Moreover, the generated images can augment the original datasets thus improve the performance in pedestrian-related tasks such as person re-identification (re-ID) and attribute prediction. Especially in person re-ID tasks our method outperforms state-of-the-art methods by a large margin.

Keywords: Attribute-aware · pedestrian image editing · data augmentation.

1 Introduction

In recent years a surge of researches on image generation have found their applications in various real-world computer vision and multimedia tasks, e.g. facial attribute editing and animation [13, 19, 2], image super-resolution [4], object detection [3], and image-to-image translation [9, 12, 10]. Among them, some recent works focus on generating pedestrian images given pose information [27, 26, 21, 16] or just from scratch [23]. The generated pedestrian images can be used to boost the performance of related learning tasks, such as person attribute prediction and person re-ID, through data augmentation of pedestrian image datasets. However, compared with general image generation, pedestrian image generation is more challenging, due to complex body configurations and poses, abundant details, variant lighting and backgrounds, etc.
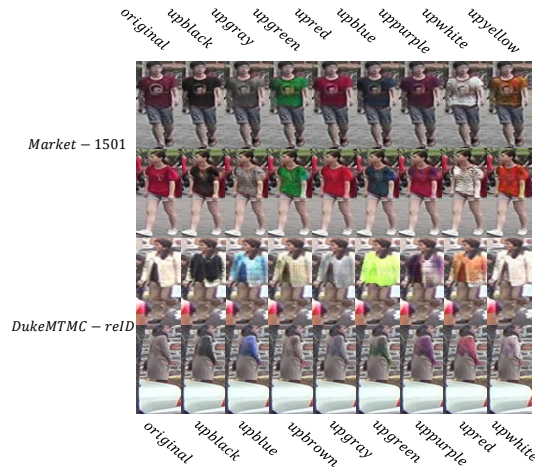
Fig. 1: The image generation examples of the APIE-Net.

The main drawbacks of the aforementioned pedestrian image generation approaches are three-folds: body distortion, inadequate visual details and large vague areas. Firstly, body distortion occurs due to ambiguous locations of different body parts. For example, the generation areas might wrongly deviate to other body parts or even the background. Secondly, while obtaining global structures of pedestrians is relatively easy for most generation models, generating rich local details is problematic due to deficient given information. Though [27, 26, 21, 16] take advantage of pose annotations in person generation, the appearance details is still inadequate. Thirdly, large vague areas appear in the generated person images because of the conflict between disparate appearances from unsettled viewpoints and (or) complex background scenarios. Neither Gaussian noise nor pose labels could provide enough information to address this problem well.

To overcome the above drawbacks, we propose to introduce visual attributes in pedestrian generation. Visual attributes contain characteristics which can greatly benefit image generation: visual attributes usually describe specific regions, e.g. "upper-body black" refers to the pixels of the upper-torso; visual attributes can incorporate affluent details, e.g. "lower-body in white skirt"; visual attributes are relatively consistent, and invariant to viewpoints and backgrounds. The pose information adopted in [27, 26, 21, 16] can be considered as a kind of visual attribute. (For concise presentation we use the term "attribute" for "visual attribute" hereafter.) In this paper, we focus on more general attribute-based pedestrian generation. With a certain specified attribute, we aim to re-generate (or edit) the image of a given pedestrian with the attribute. We thus call our method **Attribute-aware Pedestrian Image Editing** (APIE). Although at-
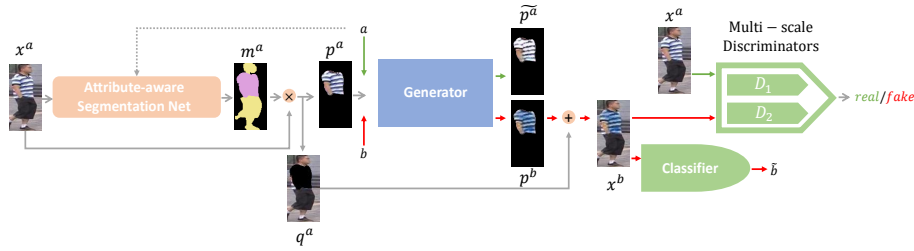
Fig. 2: The main architecture of the APIE-Net.

tributes have been used in face editing, it is the first time to study pedestrian image editing based on attributes.

Though some efforts have been made on attribute-based image generation [8–13, 29], the task is more difficult in the context of pedestrian editing due to the complexity of pedestrian images. In this paper, our proposed **APIE**-Net consists of three specific mechanisms to address the difficulties. Firstly, to avoid dramatic distortions of pedestrian appearance, we use a segmentation net before image generation to locate the region of interest, so as to centralize the generation on the target body part and preclude the effects to the other regions including background. Secondly, to generate more visual details, a series of multi-scale adversarial discriminators are adopted to capture local visual information of different granularity. Thirdly, to disentangle the attributes which blend with each other or with other image appearance, we exploit adversarial training of the latent variables like the Fader-Net [14] to remove side-effects from irrelevant attributes or image regions, e.g. for white-colored and striped cloths we want to retain the strips while switching the color.

To verify the effectiveness of our method, we utilize the attribute labels provided in [28] and generate additional pedestrian images for two datasets, Market-1501 [30] and DukeMTMC-reID [22]. Some exemplars are depicted in Fig. 1. Comparisons on the generated pedestrian images with other generation methods [14, 19, 7] show that our method brings both quantitative and qualitative improvements. On the other hand, the generated images are beneficial to other pedestrian-related visual tasks through data augmentation. Most pedestrian datasets are "attribute-imbalanced" in which some attributes have few positive exemplars while others have many. The generated pedestrian images can augment the original datasets and make them more balanced, leading to state-of-the-art performance in both person re-ID and pedestrian attribute learning tasks. This result proves the effectiveness and practicability of our method from another view.

2   Methodology

The main architecture of the APIE-Net consisting of four parts is shown in Fig.2. At the input-end an attribute-aware segmentation network is used to seg-

ment the input pedestrian image and extract the region corresponding to the given attributes. The generator consists of an encoder-decoder network with skip connections and a latent-variable discriminator used to erase contradictory attributes. The multi-scale discriminator guarantees the realness of the generated images. The attribute classifier ensures the generated images to posses the assigned attributes.

Let $\mathbf{x^a}$ denote an input pedestrian image $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$ (W, H, C denote the width, height and the number of channels respectively) with attributes $\mathbf{a} \in \{0, 1\}^l$ (l denotes the number of attribute categories). The tth element of $\mathbf{a}$ is 1 if $\mathbf{x^a}$ has the tth attribute (e.g., white cloths) and 0 otherwise. The overall target of APIE-Net is to generate a new pedestrian image $\mathbf{x^b}$ given new attributes $\mathbf{b}$ (e.g., blue cloths).

2.1   Attribute-aware Segmentation Network

In order to decide the location of the body part in which the attributes are to be changed, we resort to attribute-aware pedestrian segmentation. More specifically, we adopt a well-trained segmentation model [15, 17]. Based on the fine-grained segmentation results and user specified target, we can obtain mask $\mathbf{m^a} \in \{0, 1\}^{W \times H}$, where the pixels with value 1 denote the foreground area and 0 the background area. Note that pose information may also be utilized like [21, 16] to obtain the mask if the key points of pedestrian bodies are available. For simple and illustrative purposes, the learned mask mainly corresponds to three body parts: head, upper body, and lower body. Then, the image region to be edit can be expressed as $\mathbf{p^a} = \mathbf{x^a} \odot \mathbf{m^a}$, e.g. the upper body of a pedestrian. The other part $\mathbf{q^a} = \mathbf{x^a} \odot (\mathbf{1} - \mathbf{m^a})$ is considered as the background. $\odot$ denotes element-wise multiplication across channels.

We apply the above process to a set of pedestrian images $\mathbf{X^A}$ and obtain the editable and background regions represented as $\mathbf{P^A} = \mathbf{X^A} \odot \mathbf{M^A}$ and $\mathbf{Q^A} = \mathbf{X^A} \odot (\mathbf{1} - \mathbf{M^A})$.

2.2   The Generator

Encoder-decoder network with skip connection. The generator consists of an encoder and a decoder, denoting as $\mathbf{G}_{enc}$ and $\mathbf{G}_{dec}$ respectively, as shown in Fig.3. The encoder $\mathbf{G}_{enc}$ projects the selected image region $\mathbf{p^a}$ to the latent variable $\mathbf{z^a}$ by use of several convolution layers. Then we randomly flip one bit of the attribute vector $\mathbf{a}$ to get the new attribute vector $\mathbf{b}$, and check the conflicts to avoid irrational attribute combination (e.g. one cannot wear in white while in red). $\mathbf{b}$ will be given at test time. The concatenation of $\mathbf{b}$ and the latent variable $\mathbf{z^a}$ passes the deconvolution layers of decoder $\mathbf{G}_{dec}$, from which a new image region $\mathbf{p^b}$ is generated. Similarly, decoding the concatenation of $\mathbf{a}$ and $\mathbf{z^a}$ outputs the reconstructed original image $\tilde{\mathbf{p}}^\mathbf{a}$, where " ~ " differentiates the estimation from the ground-truth. Skip connections are adopted between corresponding convolution and deconvolution layers, making the encoder-decoder network a U-net. The above process can be expressed more formally as following: $\mathbf{Z^A} =$
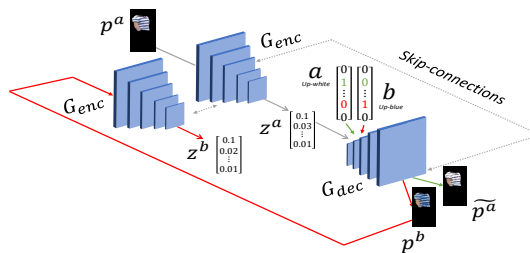
Fig. 3: The Generator with skip-connections.

$\mathbf{G}_{enc}(\mathbf{P^A})$, $\tilde{\mathbf{P^A}} = \mathbf{G}_{dec}(\mathbf{Z^A}, \mathbf{A})$, $\mathbf{P^B} = \mathbf{G}_{dec}(\mathbf{Z^A}, \mathbf{B})$. The reconstructed image regions should be as close to the original ones as possible, thus the loss function for this encoder-decoder network is defined as:

$$\mathcal{L}_{rec} = \left\| \mathbf{P^A} - \tilde{\mathbf{P^A}} \right\|^2. \tag{1}$$

Discrimination between latent variables. Some attributes of pedestrian images are entangled with each other, while some are mutually exclusive. For instance, the sex of a person is highly related to the styles of hair and clothes, while the concolorous shirt cannot be "white" and "blue" simultaneously. Therefore, when we edit an image according to the attribute $\mathbf{b}$, we have to pay attention to the attribute cooccurrence and contradiction issues. Instead of shielding the attributes $\mathbf{b}$ manually to avoid the conflict with the latent variable, a more ideal method is to automatically erase attributes from the latent variable in the encoding process. To this end, we resort to adversarial learning between latent variables. More specifically, the generated $\mathbf{p^b}$ is re-input to the encoder to get latent variable $\mathbf{z^b}$. Then, a discriminator $\mathbf{D}_Z$ is trained to capture the discrimination between latent variables $\mathbf{z^a}$ and $\mathbf{z^b}$, which is formulated as follows:

$$\mathcal{L}_{adv_{D_z}} = \mathbb{E}_{\mathbf{z^a}}[\mathbf{D}_Z(\mathbf{Z^A})] - \mathbb{E}_{\mathbf{z^b}}[\mathbf{D}_Z(\mathbf{Z^B})]. \tag{2}$$

The adversarial min-max process is

$$\min_{\mathbf{G}_{enc}} \max_{\|\mathbf{D}_Z\| \leq 1} \mathcal{L}_{adv_{D_z}}. \tag{3}$$

The min-max game between $\mathbf{G}_{enc}$ and $\mathbf{D}_Z$ learns latent variables invariant to attribute vectors. When the invariance is met, the decoder must use the attribute vector to generate image. In this way, attribute information is implicitly erased from the latent variables.

2.3   The Multi-scale Discriminators

We propose another adversarial learning process to guarantee the realness of the generated image $\mathbf{X^B} = \mathbf{P^B} + \mathbf{Q^A}$, which has the same size with the original
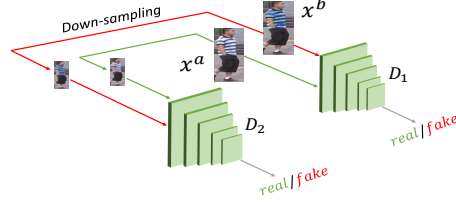
Fig. 4: The multi-scale discriminators.

image $\mathbf{X^A}$. To capture abundant local visual details, we use two discriminators in different resolutions as in [13]. As shown in Fig.4, one discriminator, denoted as $\mathbf{D}_1$, takes in the generated images as ever. The other discriminator, denoted as $\mathbf{D}_2$, receives the generated images down sampled to the half resolution. Because of the unstable training process of the original GAN model, here we use the WGAN-GP model [18]. The formulations for the discriminators are

$$\mathcal{L}_{adv_1} = \mathbb{E}_{\mathbf{x^a}}[\mathbf{D}_1(\mathbf{X^A})] - \mathbb{E}_{\mathbf{x^b}}[\mathbf{D}_1(\mathbf{X^B})] + \lambda_{GP}\mathbb{E}_{\hat{\mathbf{x}}}[\mathbf{D}_1(\hat{\mathbf{X}})], \tag{4}$$

$$\mathcal{L}_{adv_2} = \mathbb{E}_{\mathbf{x^a}}[\mathbf{D}_2(\mathbf{X^A})] - \mathbb{E}_{\mathbf{x^b}}[\mathbf{D}_2(\mathbf{X^B})] + \lambda_{GP}\mathbb{E}_{\hat{\mathbf{x}}}[\mathbf{D}_1(\hat{\mathbf{X}})], \tag{5}$$

The overall loss of multi-scale discriminators is:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv_1} + \mathcal{L}_{adv_2} \tag{6}$$

The adversarial optimization process is:

$$\min_{\mathbf{G}} \max_{\|\mathbf{D}_*\| \leq 1} \mathcal{L}_{adv}, \tag{7}$$

where $\mathbf{D}_*(\mathbf{X^B}) = \mathbf{D}_*(\mathbf{G}_{dec}(\mathbf{G}_{enc}(\mathbf{X^A} \odot \mathbf{M^A}), \mathbf{B}) + \mathbf{X^A} \odot (1 - \mathbf{M^A}))$, and $\mathbf{G}$ denotes the generator parameters in both of the encoder and the decoder. $\|\mathbf{D}_*\| \leq 1$ is the 1-Lipschitz constraint implemented by gradient penalty.

2.4   The Attribute Classifier

The APIE-Net edits pedestrian images by transferring attributes, i.e., altering one attribute and keeping the others unchanged. One way to evaluate the model capability is to investigate the attributes of the generated images. We construct a classifier for this purpose, as in [19, 14]. The classifier takes the generated image $\mathbf{x^b}$ as input and output the estimated attributes $\tilde{\mathbf{b}}$. Then, the cross-entropy loss between $\tilde{\mathbf{B}}$ and the ground-truth $\mathbf{B}$ is:

$$\mathcal{L}_{cls} = \sum_{\tilde{\mathbf{b}} \in \tilde{\mathbf{B}}, \mathbf{b} \in \mathbf{B}} -\mathbf{b}\log\tilde{\mathbf{b}} - (1 - \mathbf{b})\log(1 - \tilde{\mathbf{b}}). \tag{8}$$

It is noteworthy that the classifier share model parameters with the discriminator $\mathbf{D}_1$. The effect of attribute classification and image discrimination is reciprocal. Experiments will show the improvements on generating image details.

Table 1: The IS of attribute transformation on Market-1501

Models	up-green	up-yellow	up-purple	mean
CycleGAN	3.954	4.141	3.291	3.795
StarGAN	3.203	3.474	3.108	3.262
FaderNet	3.202	3.482	3.249	3.311
AttGAN	4.204	4.247	4.189	4.213
APIE(Ours)	**4.253**	**4.330**	**4.245**	**4.276**

Table 2: The IS of attribute transformation on DukeMTMC-reID.

Models	up-green	up-yellow	up-purple	mean
CycleGAN	3.351	4.285	1.380	3.005
StarGAN	2.246	2.551	1.600	2.132
FaderNet	3.608	3.361	2.609	3.193
AttGAN	4.218	4.014	4.500	4.244
APIE(Ours)	**4.524**	**4.222**	**4.562**	**4.436**

2.5   The Overall Loss

The generator, multi-scale discriminator and attribute classifier are trained simultaneously by optimizing the overall loss function.

$$\mathcal{L}(\mathbf{C}, \mathbf{G}_{enc}, \mathbf{G}_{dec}, \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_Z)$$
$$= \mathcal{L}_{adv} + \lambda_C \mathcal{L}_{cls} + \lambda_R \mathcal{L}_{rec} + \lambda_{D_Z} \mathcal{L}_{adv_{D_Z}}, \tag{9}$$

where λ_* denote hyper-parameters to balance the losses. Following the adversarial training process, the parameters can be obtained by the following min-max game:

$$\arg \min_{\mathbf{C}, \mathbf{G}_{enc}, \mathbf{G}_{dec}} \max_{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_Z} \mathcal{L}(\mathbf{C}, \mathbf{G}_{enc}, \mathbf{G}_{dec}, \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_Z). \tag{10}$$

3   Experiments

3.1   Datasets and implementation

To verify the effectiveness of the proposed method, three experiments are conducted, namely pedestrian images generation, person re-ID and attribute learning. All experiments are performed on the datasets **Market-1501** [30] and **DukeMTMC-reID** [22], with labels provided in [28]. These two datasets are collected for person reID task, with 12,936/16,622 images belonging to 751/702 identities for training. The images for training have size of $128 \times 64 \times 3$, and so are the generated images. The datasets are labeled with 27/23 attributes, with 8/8 attributes for upper-body clothes and 9/7 attributes for the lower-body.

The encoder and decoder of the generator network consist of 5 (de)convolution layers, so do the classifier and the multi-scale discriminator. The latent discriminator in the generator is a light-weighted net with only two fully connected layers. The initial learning rate is 0.0002 and it declines to one-tenth after every

10,000 iterations. We set $\lambda_R = 100, \lambda_C = 10, \lambda_{D_z} = 0.1$. We use Adam optimizer as in WGAN-GP model. The batch size is set as 50, and the number of training epochs is 400.

For the person re-ID experiments, we construct a baseline model by using the ResNet50 [25] as our backbone and tuning the network based on [20]. The input images are resized to 256×128, randomly cropped and horizontally flipped before training, and no dropout is adopted. The batch size is set to 32. The learning rate is initialized as 0.0003, and decreased to one-tenth after every 20 epochs. During training we augment the original datasets with the same amount of images generated by APIE-Net, which are supposed to have distinct identities. We use label smoothing regularization (LSR) as in [5] to balance the weights of the generated images. Mean average precision (mAP) and Cumulative Matching Characteristics (CMC) are evaluation metrics. As for attribute prediction, we use similar baseline model except two points: batch normalization is not adopted and multi-sigmoid loss is used instead of softmax loss. mAP is used for evaluation.

3.2   Pedestrian Image Editing

Pedestrian image editing is to generate new pedestrian images according to specified attributes. For illustration convenience, we select two identities from each dataset and change the color attributes of the clothes.

Some generated pedestrian examples by APIE-Net have been illustrated in Fig.1. For comparison, we select four commonly used researches on image editing, namely CycleGAN [7], Att-GAN [19], Fader-Net [14] and StarGAN [11]. The model parameters are set as in the paper accordingly. In the Fig.5 we demonstrate the qualitative comparison. Though these works perform well in facial attribute editing, they do poorly in the pedestrian editing scenario. The APIE-Net outperforms all of the other methods, due to its merits in image generation. First, while some attributes are changed, most details can be maintained. Second, the transformed body part can be precisely located. Third, the edited pedestrian images look natural with very few artifacts.

The quantitative comparison is listed in Table 1 and 2. Inception score (IS) is used as the evaluation metric, which measures the diversity of the generated images. The higher value means the better result. Our method achieves the highest IS in both datasets.

3.3   Data Augmentation for Person Re-ID and Attribute Prediction

We further apply APIE-Net to augment the datasets in person re-ID and attribute prediction tasks. Each augmented dataset is of double size of the original one. We construct a strong "baseline" model for each task as explained above.

We compare the performance of APIE-Net with the methods that augment the datasets for person re-ID in Table 3. The results in the upper part are the highest records reported in the corresponding papers. In the lower part, the cautiously tuned baseline model performs better than the above state-of-the-art methods. The other three methods (including ours) are built on the baseline with
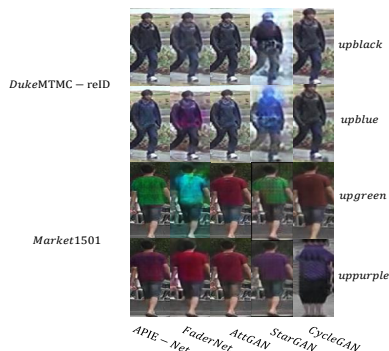
Fig. 5: The image generation examples of the **APIE**-Net.

Table 3: Comparison on person re-ID with data augmentation.

Models	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
Basel(R)+LSRO[23]	78.06	56.23	67.68	47.13
Pose-transfer[26]	87.65	68.92	78.52	56.91
IDE+CamStyle+RE[20]	89.49	71.55	78.32	57.61
baseline	92.46	78.66	83.42	68.03
pure-color filling	91.27	76.53	83.08	66.94
StarGAN	91.39	76.34	80.88	62.59
APIE(Ours)	**93.47**	**80.22**	**85.23**	**69.67**

data augmentation. However, the trivial pure-color-filling method and starGAN decrease the performance, indicating that inferior augmented images have negative effects on the task. On the other hand, our method generates images with accurate location and enough details, thus achieve the highest performance.

The comparative study on attribute learning is reported in Table 4. We compare our method with [31][28][32][33] with respect to "C.up"(*the color of the upper-body*), "C.low"(*the color of the lower-body*) and the mean accuracy, as we augment these attributes in the experiments. An average of 0.3% performance gain is achieved by APIE-Net over the baseline, and much larger gain over other methods on the Market1501 dataset. APIE performs lower than [32][33] on DukeMTMC for the lower baseline we used. The high performance attributed to the well generated images with accurate attribute information, which provides another verification to our method.

3.4   Ablation Study

Four parts are used in the APIE-Net which are alternative but useful, namely the attributes-aware segmentation net, the multi-scale discriminator ("M.D."), the latent variable discriminator ("L.D."), and the discriminator-classifier share-weight mechanism ("S.W."). The attribute-aware segmentation network and the generator-discriminator together construct the baseline set-up. "w.o. mask"

Table 4: The attribute learning experiment using the augmented data on Market-1501 and DukeMTMC-reID.

Models	Market-1501			DukeMTMC-reID		
	C.up	C.low	mean	C.up	C.low	mean
SVM[31]	-	-	-	70.90	68.50	69.70
APR[28]	73.40	69.91	71.55	72.29	41.48	57.91
Sun et al.[32]	87.50	87.20	86.35	**93.90**	91.80	**92.85**
JCM[33]	92.90	93.50	93.20	92.90	**92.10**	92.50
Baseline	95.30	94.18	94.70	88.79	82.55	85.88
Ours	**95.46**	**94.59**	**95.00**	88.95	83.19	86.26

Table 5: The IS comparison between different schemes

	Market-1501	DukeMTMC-reID	mean
baseline	4.248	4.483	4.366
w.o. mask	4.228	4.238	4.233
+M.D.	4.242	4.495	4.369
+L.D.	4.246	**4.513**	4.380
+S.W.	4.233	4.486	4.360
+M.D.+S.W.+L.D.	**4.345**	4.446	**4.396**

means the baseline without the attribute-aware segmentation net. The combination of "M.D.", "L.D." and "S.W." is the APIE-Net. To verify the effectiveness of each model part to the final performance we make the following ablation study.

As shown in Table 5, the IS values roughly increase whenever a new model part is involved. The final combination outperforms each single part. Then we make visualization of the image generation results as shown in Fig.6. The result of taking the attribute-aware segmentation net out of the baseline method is depicted in the second row, which is least satisfying for doing the least transformation.

4   Conclusion and Future Works

In this paper, we propose the problem of attribute-aware pedestrian image editing and a new model, APIE-Net, as its solution. With the aid of specified attributes, APIE-Net can generate high-quality pedestrian images, which benefit for real-world applications including person re-ID and attribute prediction. Comprehensive experiments demonstrate the effectiveness of the proposed method. In the future we will experiment on more complex attributes in pedestrian image editing.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680. (2014)
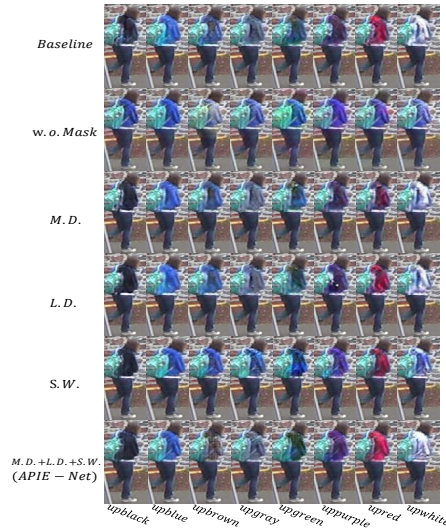
Fig. 6: Image generation results with different parts of APIE-Net.

2. Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., Moreno-Noguer, F.: GAN-imation: anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision, pp. 818–833. (2018)
3. Wang, X., Shrivastava, A., Gupta, A.: A-fast-rcnn: Hard positive generation via adversary for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2606–2615. (2017)
4. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690. (2017)
5. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826. (2016)
6. Kingma, D. P., Welling, M.: Auto-encoding Variational Bayes. In: Proceedings of the International Conference on Learning Representations. (2014)
7. Zhu, J. Y., Park, T., Isola, P., Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232. (2017)
8. Liu, M. Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in neural information processing systems, pp. 469–477. (2016)
9. Liu, M. Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation Networks. In: Advances in Neural Information Processing Systems, pp. 700–708. (2017)
10. Huang, X., Liu, M. Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision, pp. 172–189. (2018)
11. Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797. (2018)

12. Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M., Yang, M. H.: Diverse image-to-Image Translation via Disentangled Representations. In: Proceedings of the European Conference on Computer Vision, pp. 35–51. (2018)

13. Xiao, T., Hong, J., Ma, J.: ELEGANT: exchanging latent encodings with GAN for transferring multiple face attributes. In: Proceedings of the European Conference on Computer Vision, pp. 168–184. (2018)

14. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L.: Fader networks: Manipulating images by sliding attributes. In: Advances in Neural Information Processing Systems, pp. 5967–5976. (2017)

15. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: joint body parsing & pose estimation network and a new benchmark. In: IEEE transactions on pattern analysis and machine intelligence, pp. 871–885. (2019)

16. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in Neural Information Processing Systems, pp. 406–416. (2017)

17. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 932–940. (2017)

18. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems, pp. 5767–5777. (2017)

19. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: Arbitrary facial attribute editing: only change what you want. arXiv preprint arXiv:1711.10678 (2017)

20. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5157–5166. (2018)

21. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 99–108. (2018)

22. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision, pp. 17–35. (2016)

23. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762. (2017)

24. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 994–1003. (2018)

25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. (2016)

26. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pp. 4099–4108. (2018)

27. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., ... Xue, X.: Pose-normalized image generation for person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 650–667. (2017)

28. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220. (2017)

29. Chen, X., Xu, C., Yang, X., Tao, D.: Attention-gan for object transfiguration in wild images. In: Proceedings of the European Conference on Computer Vision, pp. 164–180. (2018)
30. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124. (2015)
31. Kurnianggoro, L., Jo, K. H.: Identification of pedestrian attributes using deep network. In: IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society, pp. 8503–8507. (2017)
32. Sun, C., Jiang, N., Zhang, L., Wang, Y., Wu, W., Zhou, Z.: Unified framework for joint attribute classification and person re-identification. In: International Conference on Artificial Neural Networks, pp. 637–647. (2018)
33. Liu, H., Wu, J., Jiang, J., Qi, M., Bo, R. : Sequence-based person attribute recognition with joint ctc-attention model. arXiv preprint arXiv:1811.08115. (2018)