Received: 29 July 2021 | Revised: 29 September 2021 | Accepted: 21 October 2021

IET Computer Vision

DOI: 10.1049/cvi2.12081

ORIGINAL RESEARCH PAPER

HQA-Trans: An end-to-end high-quality-awareness image translation framework for unsupervised cross-domain pedestrian detection

Gelin Shen[1] | Yang Yu[1] | Zhi-Ri Tang[2] | Haoqiang Chen[1] | Zongtan Zhou[1]

[1]College of Intelligence Science and Technology, National University of Defense Technology, Changsha, Hunan, China

[2]School of Physics and Technology, Wuhan University, Wuhan, Hubei, China

Correspondence

Yang Yu, College of Intelligence Science and Technology, National University of Defense Technology, Changsha, Hunan, China.
Email: yuyangnudt@hotmail.com

Zhi-Ri Tang, School of Microelectronics, Wuhan University, Wuhan, Hubei, China.
Email: gerintang@163.com

Funding information

National Natural Science Foundation of China, Grant/Award Number: 62006239; joint Funds of National Natural Science Fundation of China, Grant/Award Number: U19A2083

Abstract

Unsupervised cross-domain pedestrian detection has attracted attention in recent years. Although some works adopted unsupervised image translation frameworks to generate an intermediate domain to narrow the gap between source and target domains, the images in the intermediate domain tend to be distorted due to the instability of the generation network. In this work, we propose a new framework to improve the image quality of the generated intermediate domain via an end-to-end translation framework. First, an image quality assessment index is adopted and adjusted appropriately. The part that controls the image quality is kept, and the part that adversely affects the domain style translation is discarded. Secondly, the adjusted image quality assessment index is integrated into the unsupervised image translation framework, where a new loss with the index's weight is proposed. An end-to-end high-quality-awareness image translation framework is constructed to generate a high-quality intermediate domain directly through this process. Finally, the intermediate domain with high-quality images is applied for cross-domain pedestrian detection. Experimental results on benchmark datasets show that the proposed framework can effectively improve unsupervised cross-domain pedestrian detection performance. Compared with some state-of-the-art works, the proposed framework can also achieve superior performance under miss rate metrics.

1 | INTRODUCTION

Pedestrian detection occupies an essential position in the research of computer vision, which is mainly to classify and locate pedestrians in the given images or videos. In practical applications, because of the high cost and the difficulty of obtaining a large number of labels, researchers have turned their attention to unsupervised pedestrian detection [1]. However, for the unlabelled target domain, the detector's performance is usually severely reduced due to the difference in scene distribution between the source and the target domains. In other words, the generalisation ability to detect directly from the source domain to the target domain is unsatisfactory. For this reason, an unsupervised image translation framework is introduced into unsupervised cross-domain pedestrian detection to generate an intermediate domain with a scene style close to the target domain [2–5]. It can reduce the difference between domains, improve the generalisation ability between domains, and enhance the self-adaptability of the domain, thereby effectively improving the performance of cross-domain detection.

Using the unsupervised image translation framework, the source image is translated into a target-like source domain with both the contents of the source domain and the style characteristics of the target domain. It is also called the intermediate domain, transforming the original cross-domain pedestrian detection problem into a cross-domain task between the intermediate and target domains. Since the similarity of the scene distribution between the intermediate and the target domains is higher than that between the source and the target ones, generally speaking, applying an unsupervised image translation framework can effectively improve the cross-domain performance. However, due to the instability of the generation network, the images in the intermediate domain usually suffer from different degrees of distortions, and some low-quality images are also generated, which undoubtedly hurts the cross-domain performance.

In this work, a new framework, named HQA-Trans, is proposed, which incorporates an image quality assessment index [6] into the unsupervised image translation framework for unsupervised cross-domain pedestrian detection. The proposed end-to-end translation framework can effectively control the generated image quality in the intermediate domain and avoid image distortions. Our contributions can be summarised as follows:

(1) The image quality assessment index is integrated into the unsupervised image translation framework. The index is adjusted appropriately, which effectively retains the contents and structural characteristics of the original images and does not affect the style conversion of the domain. It can help to improve the image quality, where only a few distortions can be found in the generated domain. In other words, the proposed method can effectively avoid the significant distortions of the images caused by the instability of the original generation network
(2) The integration of the image quality assessment index into the unsupervised image translation framework ensures that the image quality can be under control during image generation. The end-to-end translation framework avoids manual screening and replacement of low-quality images in the dataset before and after image generation, which has greatly reduced manpower consumption and improved the efficiency of obtaining a high-quality dataset
(3) Experimental results on two benchmark pedestrian detection datasets show that the datasets generated by the proposed framework can effectively improve the performance of cross-domain pedestrian detection. Compared with other state-of-the-art works, the proposed framework also shows its superiority

2 | RELATED WORK

2.1 | Image translation

Image translation refers to mapping images in the source domain to the target domain, including many practical applications like style conversion. The original image translation method is mainly a supervised image translation model, which requires paired images during the training process. In practical applications, many paired datasets are difficult to obtain and costly, limiting the development of supervised image translation methods. As a result, unsupervised image translation methods are very popular, and some unsupervised image translation frameworks, including CycleGAN [7] and UNIT [2] have been proposed one after another. CycleGAN designed cycle consistency (CC) to achieve image translation, while UNIT introduced a shared latent space based on CycleGAN to achieve image translation. However, none of the above research work can explain the multimodality of image translation. Then, Multimodal Unsupervised Image-to-Image Translation [8] and Diverse Image-to-Image Translation Via Disentangled Representations [3] are proposed. They encode the content and attributes of global images, which realize image translation by exchanging their attribute codes. The above two methods are both methods of directly performing image-level translation without considering the object instance. Although they can achieve certain results, they often generate image results that lack authenticity for complex scenes. Therefore, InstaGAN [9] was proposed to solve the instance-level translation problem. Furthermore, Detection-based Unsupervised Image Translation [4] was proposed, which extracted representations both in image and instance level, and then generated translated images by merging features into a public representation.

2.2 | Object detection

Predefined anchors of different sizes and proportions are necessary for traditional object detection to obtain the width and height of the target instances. Anchor-based detectors can be classified into two main types: one-stage and two-stage. Some studies of two-stage detectors such as Object Detection via Region-based Fully Convolutional Networks [10] can generate auxiliary bounding boxes for the targets and then give the classification results. Concerning one-stage detectors, YOLO [11] is one of the representative studies, which does not support any auxiliary bounding boxes for the targets, thereby improving the operating efficiency of the detection system.

Besides the mentioned anchor-based detectors above, a few anchor-free detectors have also been proposed and developed recently, such as CenterNet [12]. Besides, Center and Scale Prediction (CSP) [13] also put forward a detection head, which can predict the centre point and scale from the feature map extracted by the deep convolutional neural network (CNN) directly.

2.3 | Unsupervised pedestrian detection

A loose covariate offset hypothesis for UDA, which proposed a general generative modelling framework, was put forward by T. Adel et al. [14] in 2017. In 2018, Y. Chen et al. [15] modified Faster R-CNN into a domain adaptive version. In 2019, an automatic adaptation of the detection network to the target domain was proposed by A. Roychowdhury et al. [16]. Generative Adversarial Networks (GANs) have recently been used for image-level domain translation. GAN was applied to bridge the source domain and the target domain by generating a new domain by H. K. Hsu et al. [5]. At the same time, GAN was also used to generate new pedestrians from training source data in another study [1], giving more complicated and plentiful features in training data.

2.4 | Image quality assessment

There are two main types of image quality assessment methods: subjective image quality assessment and objective image quality assessment. In the study of subjective image quality assessment methods, not only subjective image quality assessment experiments such as absolute category rating and pair comparison [17] are included; it also includes Waterloo [18] and other image quality assessment databases. For objective image quality assessment methods, it can be divided into traditional methods and deep-based methods. In traditional methods, there are full reference image quality assessment methods such as structural similarity index (SSIM) [6], reduced reference image quality assessment methods, and no reference image quality assessment methods such as Natural Image Quality Evaluator [19]. Deep-based image quality assessment methods include Score-based Model [20], Rank-based Model [21], and Multi-task Model [22].

3 | METHOD

The difficulty of cross-domain pedestrian detection is the huge difference in the scene distribution between the source and the target domains. Therefore, building a bridge between the source and the target domains is necessary to obtain an intermediate domain, which can help reduce the difference in the scene distributions and improve detection performance. This task can be accomplished by UNIT [2],

$$\begin{aligned} \min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} \; & L_{VAE_1}(E_1,G_1)+L_{GAN_1}(E_1,G_1,D_1)+L_{CC_1}(E_1,G_1,E_2,G_2) \\ & L_{VAE_2}(E_2,G_2)+L_{GAN_2}(E_2,G_2,D_2)+L_{CC_2}(E_2,G_2,E_1,G_1) \end{aligned} \tag{1}$$

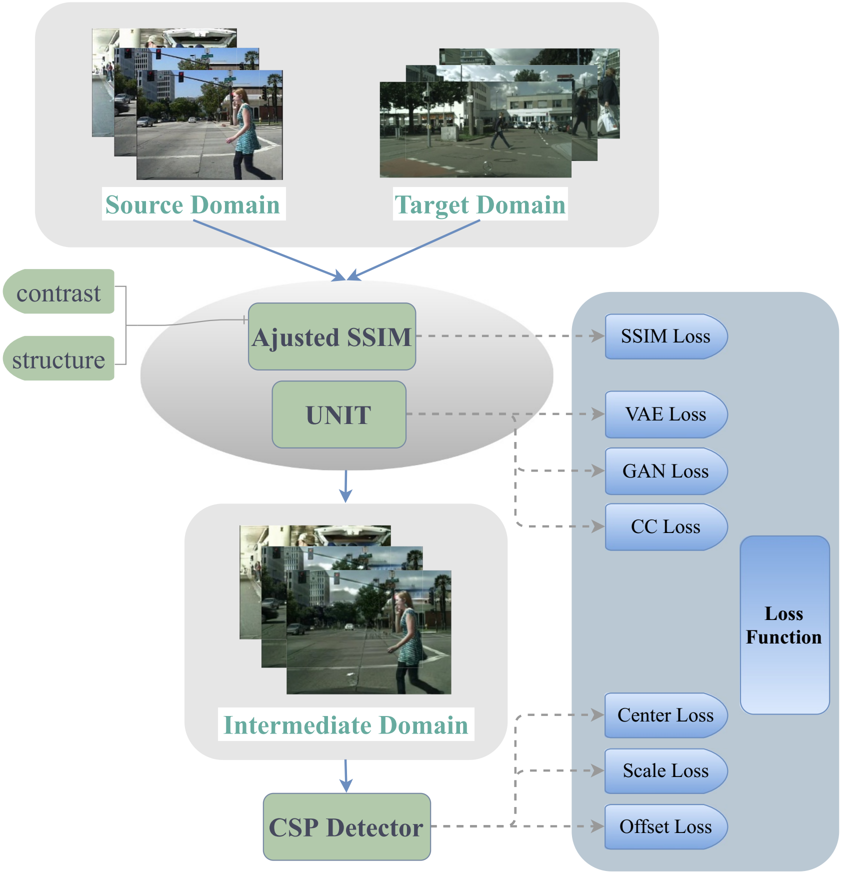

FIGURE 1 An overview of the HQA-Trans framework. CC, cycle consistency; GAN, generative adversarial network; SSIM, structural similarity index; and VAE, variational autoencoder

an unsupervised image translation framework. However, the generation network is highly unstable. The intermediate domain tends to have a large degree of distortion and even some low-quality images, which is not conducive to cross-domain detection performance. This paper proposes integrating the SSIM [6] index of image quality assessment into the unsupervised image translation framework to realize the domain style conversion of the images while ensuring that the contents and structure characteristics of the intermediate domain images are as close as possible to the source domain images. Finally, CSP is selected as the detection backbone due to its superior performance. An overview of the proposed framework, HQA-Trans, is shown in Figure 1. It shows that incorporating the adjusted SSIM index into the UNIT framework can directly generate high-quality images in the intermediate domain, which have both the contents of the source domain and the style characteristics of the target domain.

3.1 | Unsupervised image translation framework incorporating image quality assessment

3.1.1 | Scene translation

UNIT [2], as an unsupervised image translation framework, could combine the content characteristics of the source images and the style characteristics of the target images to obtain a target-like source domain, which is also called the intermediate domain. A shared latent space is proposed in UNIT, which contains two Variational Auto Encoder-Generative Adversarial Networks with a cyclic consistency loss. In this way, UNIT could translate images from one domain to another and perform a two-way translation at the same time to obtain two intermediate domains. The illustration of the UNIT framework is shown in Figure 2. In the figure, taking the intermediate domain A → B as an example, the generated domain has a similar scene distribution to domain B while the locations, sizes, and shapes of the pedestrians are the same as the source images in domain A. The intermediate domain B → A has the opposite translation result, which lays a solid foundation for the training process using the original labels of domain A.

3.1.2 | Image quality assessment

Since the no-reference image quality assessment index has too much randomness and tends to be difficult to control when integrated into the UNIT framework, the reference image quality assessment index is naturally considered. The source domain images will be used as the reference images. One problem is that the generated images may be too close to the source domain images to achieve domain style conversion.

This problem can be solved by selecting an SSIM [6] as an image quality assessment index. SSIM contains three indicators of luminance, contrast, and structure. To avoid the domain style conversion being affected, SSIM is adjusted to retain the structural and contrast indicators and ignore the luminance indicator. This adjustment will have two benefits. On the one hand, the domain style conversion of the generated images could still be controlled by the UNIT network. On the other hand, SSIM could control the similarity between the generated images and the source images in terms of content structure characteristics. As for the method of how to combine the UNIT framework and SSIM to work in the image generation process, we will discuss it in detail in Section 3.1.3.

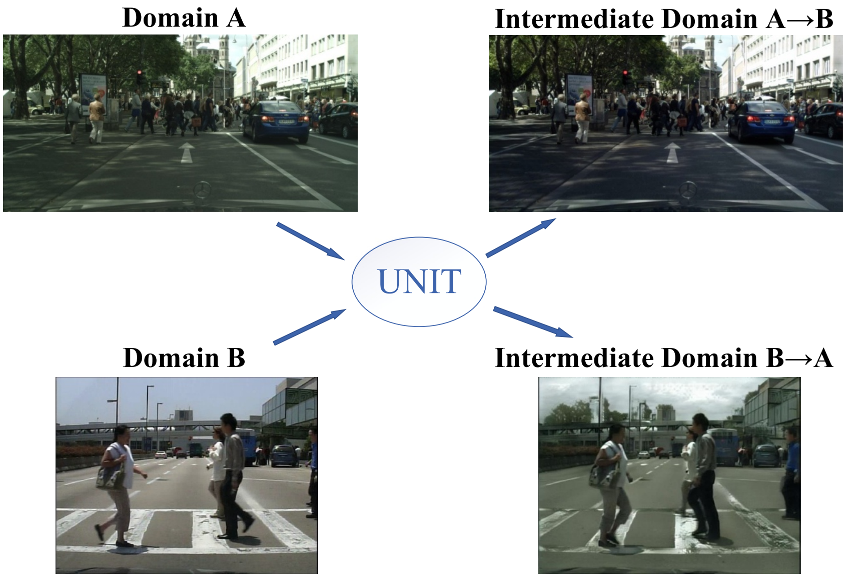

FIGURE 2 Intermediate domains generated by a two-way translation through the UNIT

The SSIM formula is based on three comparison measures between samples x and y: luminance, contrast, and structure as shown in the following formulas:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$
$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{2}$$
$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

where $c_3 = c_2/2$. μ_x and μ_y denote the mean values of x and y, respectively. σ_x^2 and σ_y^2 denote the variances of x and y, respectively. σ_{xy} is the covariance of x and y. Furthermore, two constants $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ are used to avoid division by zero. L refers to the range of pixel values, namely $(2^B - 1)$. k_1 and k_2 are the default values, which are 0.01 and 0.03, respectively.

Hence, we have:

$$SSIM(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma] \tag{3}$$

Set α, β, and γ to 1, you can get:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4}$$

In the above, an N*N window is taken from the picture. Then the window is continuously sliding for calculation. Finally, the average value is taken as the global SSIM.

In addition, the range of SSIM is $[-1, 1]$. The larger the value, the smaller the gap between the output image and the reference image, that is, the better the output image quality. When the two images are the same, the SSIM value is 1. In the three comparison measures of SSIM, the mean is used to estimate the luminance, the standard deviation is used to estimate the contrast, and the covariance is used to measure the similarity of the structure.

While we implement domain style conversion for the generated images, SSIM is selected to control the quality of the generated images. SSIM needs to be adjusted to ensure that the reference image quality assessment index will not affect the domain style conversion of the generated network. Here, the luminance index is removed where α is set to 0 and the other values remain unchanged. Thus we have:

$$SSIM(x,y) = \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{5}$$

3.1.3 | Loss function

In the scene translation module corresponding to the image reconstruction streams, the image translation streams, and the cycle-reconstruction streams, the loss function in turn includes variational autoencoder (VAE) loss, GANs loss, CC loss etc. The loss function in the translation module is as follows (see Equation (1)).

SSIM is chosen as the image quality assessment index and adjusted to score the similarity of the content structure characteristics between the generated images and the reference source images. The higher similarity means that the generated images are less likely to be distorted. Therefore, the adjusted SSIM index is incorporated into the total loss function of the UNIT framework as a new loss and its corresponding weight should be adjusted reasonably to control the quality of images during generation. The main idea is to make the generated images subject to SSIM Loss in addition to the original GAN Loss, CC Loss, and other losses. In this way, SSIM Loss can act as a constraint on the image quality during the image generation process, thus reducing the occurrence of distortion in the generated images and ultimately improving the quality of the generated intermediate domain images.

3.2 | Cross-domain pedestrian detection

The high-quality intermediate domain dataset generated by the unsupervised image translation framework integrated with the image quality assessment index (HQA-Trans) is used for bidirectional cross-domain pedestrian detection. The final cross-domain detection miss rate is used to evaluate the effectiveness of the proposed HQA-Trans method. The selected detector for pedestrian detection is the CSP [13] detector, mainly based on the centre and scale prediction. A detection head is proposed in the CSP detector and its loss functions are as follows:

$$L = \lambda_c L_{center} + \lambda_s L_{scale} + \lambda_o L_{offset} \tag{6}$$

where λ_c, λ_s, and λ_o are the corresponding weights of the centre, scale, and offset loss of CSP and are set as 0.01, 1, and 0.1, respectively.

4 | EXPERIMENTS

4.1 | Experiment settings

4.1.1 | Implement details

In order to provide sufficient computing power for deep neural network training, the computing device is equipped with RTX 2080Ti GPU. In the first part of the work, which is to integrate the SSIM index into the scene translation framework, the randomly cropped size of the input images is set to 256 × 256 to adapt to the configuration of the computing device. In addition, the initial learning rate is set to $1{*}e^{-4}$ and the batch size is set to 1. Second, ResNet-50 [23] is selected as the backbone of the detection network, and the batch size is set to 2. In the training of the City-persons dataset, the image size is adjusted to 240 × 480 and the initial learning rate is set to $2{*}e^{-4}$. In the training of the Caltech dataset, the image size is adjusted to 240 × 320 and the learning rate is set to $1{*}e^{-4}$. In addition, due to the different features of datasets, the weights of SSIM loss corresponding to different conversion directions will be different to successfully realize the high-quality domain style conversion of the datasets. Here, Caltech to Citypersons and Citypersons to Caltech are set to 10 and 2, respectively.

4.1.2 | Datasets

To evaluate the experimental performance of cross-domain pedestrian detection, two benchmark pedestrian detection datasets are selected, namely Citypersons [24] and Caltech [25]. For the Citypersons dataset, it is obtained from the Cityscapes [26] dataset, which is a high-quality pedestrian detection dataset with bounding box annotations. The Citypersons dataset has a more complex environment and more pedestrians. For the Caltech dataset, as a relatively large-scale pedestrian database at present, it was taken by In-vehicle Cameras that contain about 10 h of 640 × 480 30 Hz video, including 350,000 rectangular boxes and 2300 pedestrians. Furthermore, the situations of the time correspondence between the rectangular boxes and their occlusion are also marked. The relevant information of the Citypersons and Caltech datasets is shown in Table 1.

TABLE 1 Details of Caltech and Citypersons datasets

Dataset	Caltech [25]	Citypersons [24]
Number of training image	42,782	2975
Number of test image	4024	500
Resolution ratio	640 × 480	2048 × 1024
Country	1	3
City	1	18
Season	1	3
Unique person	1,273	19,654

4.2 | Experimental results and analysis

The generated domain using the proposed HQA-Trans framework and the generated domain using only the UNIT framework are compared qualitatively and quantitatively. In the qualitative comparison, it is mainly to display and compare the images in the generated domain. In the quantitative compari-

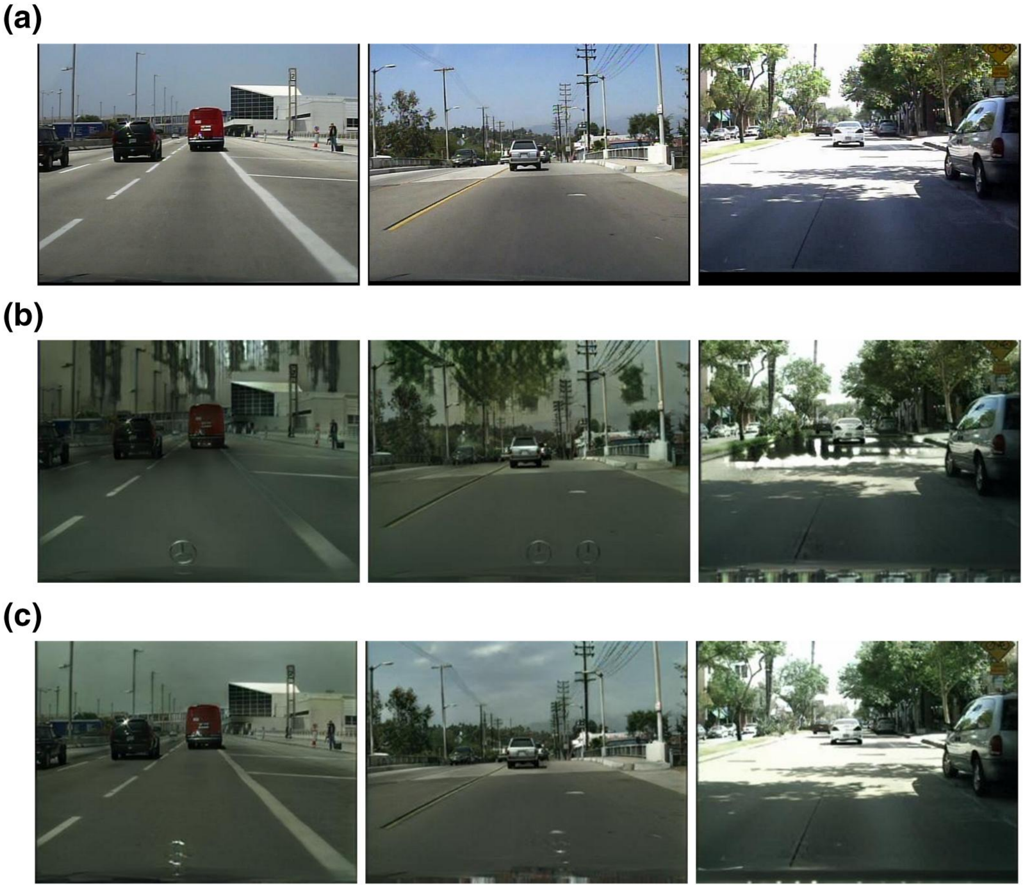

FIGURE 3 Some samples of the Caltech dataset in different scenarios

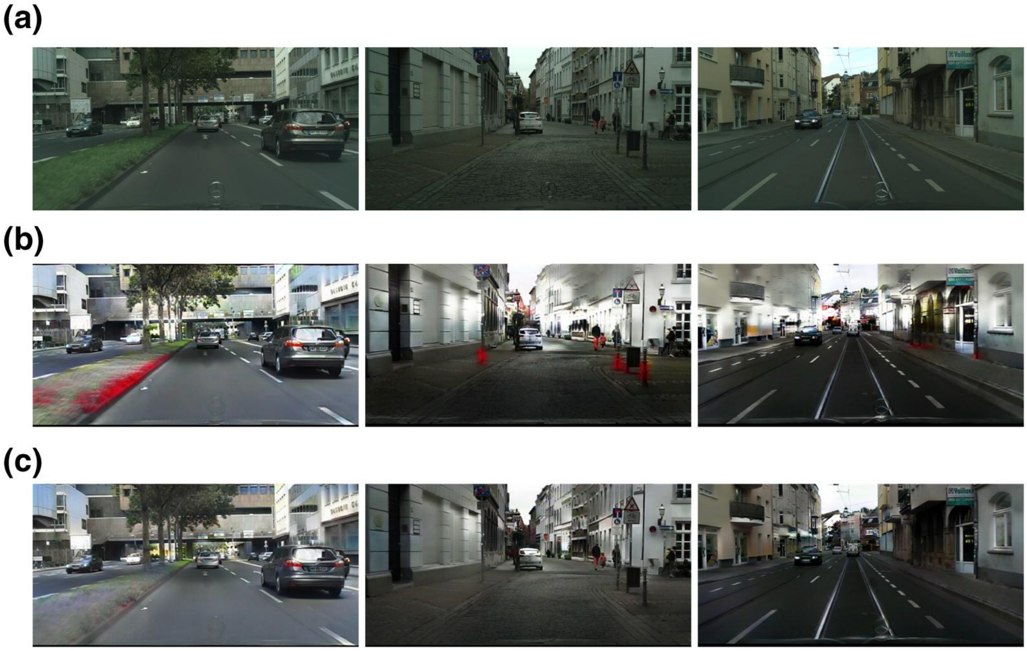

FIGURE 4 Some samples of the Citypersons dataset in different scenarios

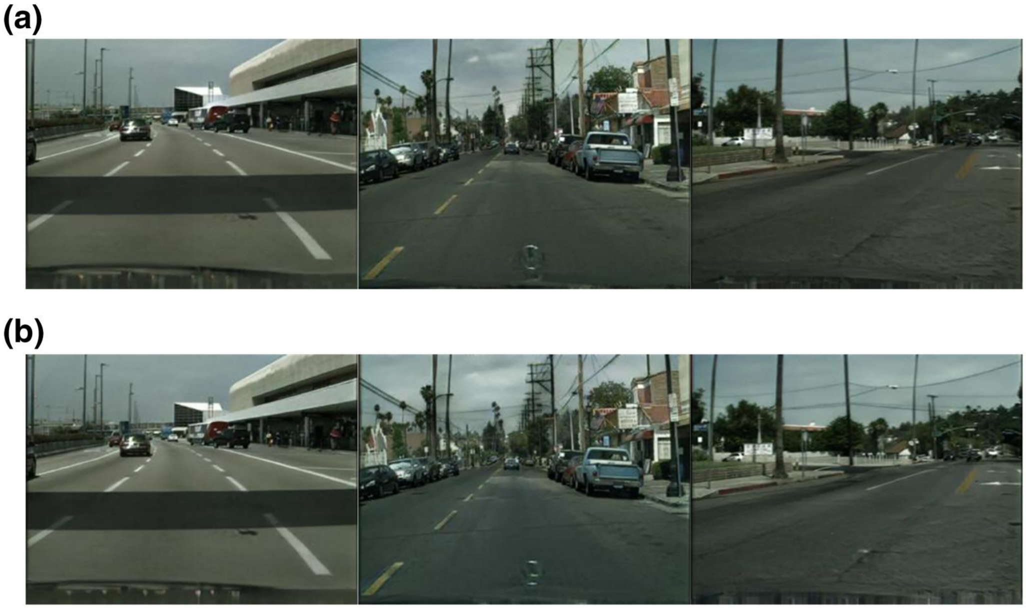

FIGURE 5 Some samples in the Caltech dataset, which successfully achieve domain style conversion and obtain the desired effect when the structural similarity index loss weight is set to 10. (a) The desired effect. (b) The successful conversion effect

son, the two generated domains will be scored through the image quality assessment indexes peak signal-to-noise ratio (PSNR) and SSIM to obtain the distributions of the scores and to evaluate and compare the quality of generation, respectively.

The first step is to qualitatively compare the two generated domains using the HQA-Trans framework and the UNIT framework as shown in Figures 3 and 4. Among them, some samples of the original dataset (Original Dataset), the intermediate domain generated by the UNIT framework (UNIT-Only Dataset), and the intermediate domain generated by the HQA-Trans framework (HQA-Trans Dataset) of Caltech and Citypersons are displayed in the figures. In Figure 3, the sky in the first and second images of UNIT-Only Dataset is seriously distorted and mixed together with the telegraph poles on the roadside while that of HQA-Trans Dataset does not appear to have any distortion, and which can follow the clearness of the images in the Original Dataset to generate a cloudy sky. The third image in the UNIT-Only Dataset shows the distortions of the road surface, which are similar to the

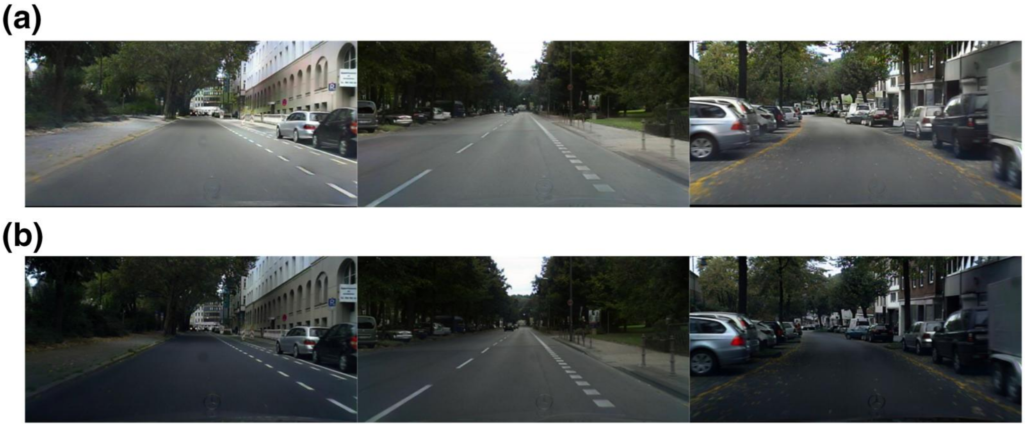

FIGURE 6 Some samples in the Citypersons dataset, which does not successfully achieve domain style conversion and does not obtain the desired effect when the SSIM loss weight is set to 10. (a) The desired effect. (b) The failed conversion effect

water and the reflection of trees, but there is no distortion in the HQA-Trans Dataset. In Figure 4, in the UNIT-Only Dataset, the roadsides, pillars etc. turned red in the first and second images, and the second and third images have problems on the generated buildings, which are extremely vague, doors and windows cannot be distinguished clearly, and there are fake pedestrians in the doors of the buildings on the left in the third image. However, no obvious distortion is seen in the images in the HQA-Trans Dataset. Figures 3 and 4 illustrate that both UNIT-Only Dataset and HQA-Trans Dataset can implement the domain style conversion. However, the images in the UNIT-Only Dataset often suffer from various types of distortions, among which some are seriously distorted. The images in the HQA-Trans Dataset nearly perfectly retain the contents and structure characteristics of the corresponding source images, which have a much stronger sense of reality. Above these show that the integration of the SSIM plays an important role in the image generation process. By using the HQA-Trans method, the images can effectively maintain the contents and structural characteristics of the original images during the generation process instead of image distortion, thereby effectively improving the image quality of the generated datasets.

In addition, as mentioned in Section 4.1.1, the weights of SSIM loss are eventually set to 10 and 2. However, at first, both weight values were set to 10 based on the magnitude judgement only. The result is as shown in Figures 5 and 6, which can be seen as failure examples of the HQA-Trans method. The dataset generated in the Caltech to Citypersons direction achieved the desired effect. However, the dataset generated in the Citypersons to Caltech direction did not succeed in realizing domain style conversion, and the overall brightness was darker than expected, with style closer to the source domain. Due to the differences in the complexity of the content structure of different datasets, there is a difference in the weight value between the two directions. We consider that the weight of the Citypersons to Caltech direction is not reasonable. When the weight is set to 10, it leads to too much constraint of SSIM loss in the generation of the intermediate domain dataset and affects the domain style conversion process. When the weight is reduced to 2, the ideal effect could be obtained. Therefore, to get the perfect generated images, the two weight values of SSIM loss need to be reasonably set.

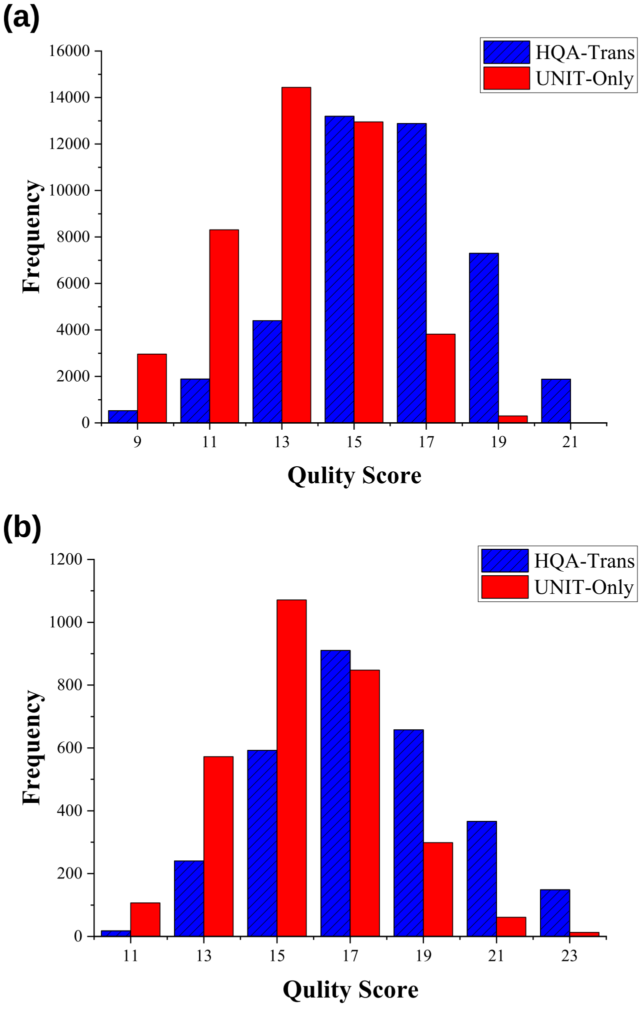

FIGURE 7 Histogram of the peak signal-to-noise ratio (PSNR) quality score of the datasets. (a) Caltech dataset. (b) Citypersons dataset

The second step is to quantitatively compare the HQA-Trans and UNIT-Only methods. As shown in Figures 7 and 8, the scoring and comparison results of Caltech Dataset and Citypersons Dataset by image quality assessment indexes PSNR and SSIM are, respectively, displayed. In the figures, the abscissa point represents the interval centre of the corresponding quality score segment. The rectangular box represents the frequency of the quality score within the range of ±1 or ±0.05 of the abscissa point. From the comparison results of HQA-Trans and UNIT-Only in the figures, it can be seen that the frequency of UNIT-Only is much higher than that of HQA-Trans in lower quality scores. As the quality score increases, the frequency of HQA-Trans starts to be higher than UNIT-Only. With regard to higher quality scores, the frequency of HQA-Trans is much higher than that of UNIT-Only. The higher the PSNR and SSIM scores, the higher the image quality; therefore, it can be inferred that the proposed HQA-Trans framework can efficiently improve the image quality of the generated datasets .

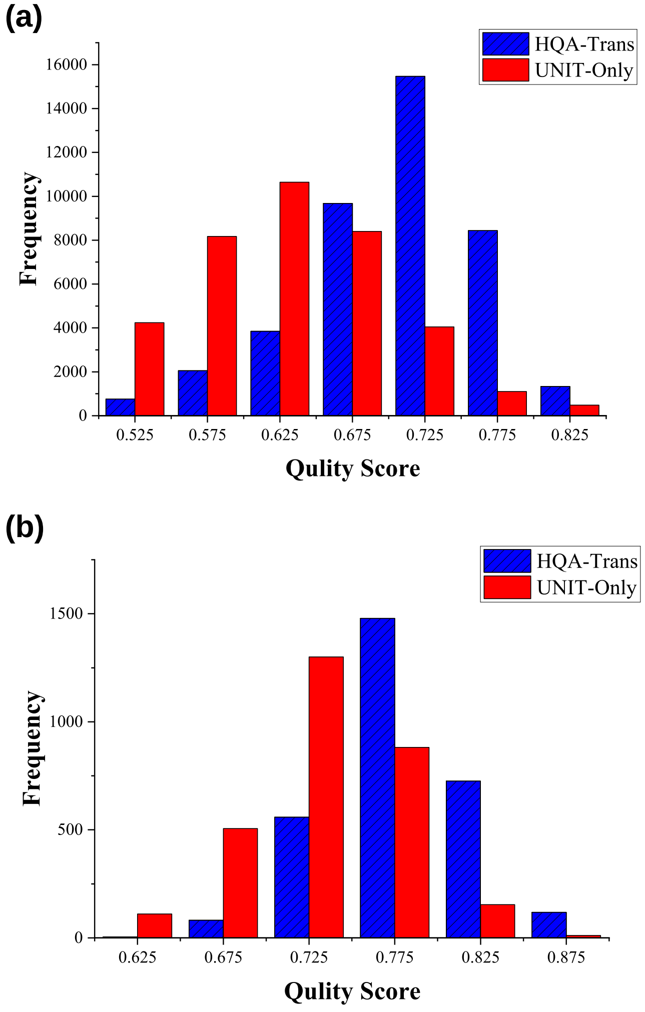

FIGURE 8 Histogram of the structural similarity index quality score of the datasets. (a) Caltech dataset. (b) Citypersons dataset

Besides, the PSNR and SSIM image quality scores in each scene of the Citypersons dataset are recorded in Tables 2 and 3 , respectively. The name of each scene is expressed in abbreviated form for convenient tabulation. The best results are shown in bold. From the two tables, it can be seen that the image quality score of HQA-Trans is higher than that of UNIT-Only in each scene whether by PSNR or by SSIM. Therefore, not only can it be obtained from Figures 7 and 8 that the overall image quality of the Citypersons dataset under HQA-Trans is better than UNIT-Only, but it can also be obtained from Tables 2 and 3 that the image quality in each scene of the Citypersons dataset under HQA-Trans is also better than UNIT-Only.

TABLE 4 Cross-domain pedestrian detection performance comparisons of Original Dataset, UNIT-Only Dataset, and HQA-Trans Dataset from Citypersons to Caltech

Training datasets	MR^{-2} (%)
Original	19.13
UNIT-only	13.87
HQA-Trans (Ours)	**12.01**
Oracle	8.04

Note: The best results are shown in bold.

TABLE 2 The PSNR quality scores in each scene of the Citypersons dataset

method	Aac	Boc	Bre	Col	dar	dus	Erf	ham	han	jen	kre	mon	stra	stu	tub	ulm	wei	zur
UNIT-only	13.43	14.90	15.18	15.95	17.86	13.33	13.22	15.89	15.89	14.78	16.70	16.66	16.08	16.59	16.02	16.77	17.22	16.62
HQA-Trans	**16.90**	**16.12**	**17.33**	**18.56**	**19.78**	**17.71**	**17.04**	**16.69**	**16.76**	**17.18**	**18.09**	**17.98**	**17.17**	**18.21**	**17.31**	**19.39**	**19.02**	**18.50**

Note: The best results are shown in bold.

Abbreviations: PSNR, peak signal-to-noise ratio.

TABLE 3 The structural similarity index quality scores in each scene of the Citypersons dataset

method	Aac	Boc	Bre	Col	dar	dus	erf	ham	han	jen	kre	mon	str	stu	tub	ulm	wei	zur
UNIT-only	0.698	0.710	0.730	0.741	0.761	0.690	0.709	0.718	0.731	0.731	0.755	0.760	0.735	0.756	0.748	0.768	0.748	0.747
HQA-Trans	**0.775**	**0.760**	**0.779**	**0.788**	**0.787**	**0.778**	**0.784**	**0.753**	**0.768**	**0.788**	**0.796**	**0.797**	**0.772**	**0.785**	**0.779**	**0.809**	**0.781**	**0.783**

Note: The best results are shown in bold.

Furthermore, CSP is selected as the detection framework. The Original Dataset, UNIT-Only Data set, and HQA-Trans Dataset obtained above are used for cross-domain pedestrian detection. In Tables 4 and 5, the unsupervised detection results are marked as 'Original', 'UNIT-Only', and 'HQA-Trans'. In addition, the fully supervised detection result is marked as 'Oracle', which can be used as a reference for unsupervised detection performance. The best unsupervised detection results are shown in bold, and the MR^{-2} measure means the miss rate of the detection. So the lower the MR^{-2}, the better the detection result. It can be seen from Tables 4 and 5 that HQA-Trans has further improved the performance of cross-domain pedestrian detection when compared with UNIT-only. Specifically, for Citypersons to Caltech, HQA-Trans can succeed UNIT-Only by 1.86% and Original by about 7.12%. For Caltech to Citypersons, HQA-Trans can lead UNIT-Only by 3.11% and Original by about 4.66%.

TABLE 5 Cross-domain pedestrian detection performance comparisons of Original Dataset, UNIT-Only Dataset, and HQA-Trans Dataset from Caltech to Citypersons

Training datasets	MR^{-2} (%)
Original	55.94
UNIT-only	54.39
HQA-Trans (Ours)	**51.28**
Oracle	23.28

Note: The best results are shown in bold.

In addition, the performance of the above three is qualitatively compared via the pedestrian bounding box, which is shown in Figure 9. The green boxes indicate that the pedestrians are correctly detected, the yellow boxes indicate that the detected results are redundant, and the red boxes indicate the pedestrians that are missed. For the first sample image in the figure, HQA-Trans has four green boxes, UNIT-Only and Original both have three green boxes, one yellow box and one red box. In the second sample image, all three have two green boxes and three red boxes, but HQA-Trans only has two yellow boxes while UNIT-Only and Original have five yellow boxes. Thus, it can be seen that the proposed HQA-Trans has the best accurate cross-domain detection performance.

4.3 | Comparisons with other work

To compare the cross-domain detection results of Citypersons to Caltech with state-of-the-art work, such as Adapted FasterRCNN [24], ACF [27], PRNet [28], ALFNet [29], and

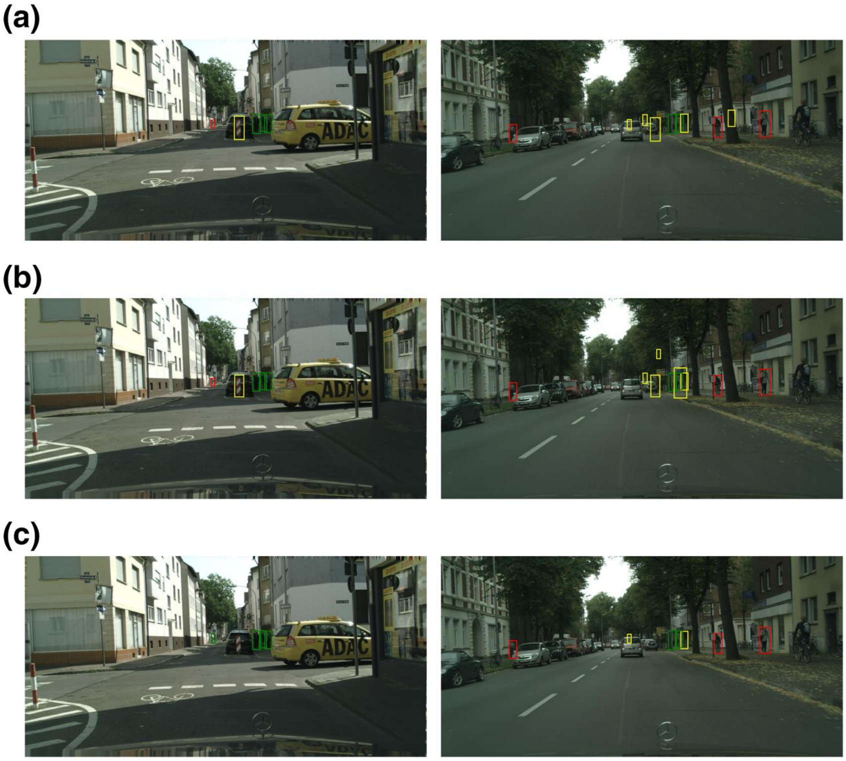

FIGURE 9 Bounding box diagram for pedestrian detection. (a) Original. (b) UNIT-Only. (c) HQA-Trans

TABLE 6 Unsupervised cross-domain detection performance and comparisons with state-of-the-art work from Citypersons to Caltech

Method	Adapted FasterRCNN [24]	ACF [27]	PRNet [28]	ALFNet [29]	APGAN [1]	HQA-Trans (Ours)	Oracle
MR^{-2} (%)	21.18	51.28	18.3	25.0	20.5	**12.01**	8.04

Note: The best results are shown in bold.

TABLE 7 Unsupervised cross-domain detection performance and comparisons with state-of-the-art work from Caltech to Citypersons

Method	ACF [27]	CSP [13]	HQA-Trans (Ours)	Oracle
MR^{-2} (%)	72.89	55.94	**51.28**	23.28

Note: The best results are shown in bold.

Abbreviations: CSP, Center and Scale Prediction.

APGAN [1], the results are shown in Table 6. Besides, ACF [27] and CSP [13] are compared with HQA-Trans to show the cross-domain detection performance of Caltech to Citypersons as shown in Table 7. In Tables 6 and 7, 'Oracle' refers to the fully supervised detection result, which can be used as a reference for unsupervised detection performance and the rest are all unsupervised methods. The best unsupervised detection results are shown in bold.

The ACF framework achieves target detection by extracting features from the approximate complex images of adjacent scales. The ALFNet architecture, constructed by four levels of feature maps for detecting objects with different sizes, aims to learn efficient single-stage pedestrian detectors. However, the generalisation ability of ACF and ALFNet is not good enough, which limits the applications on many other detection tasks. The Adapted FasterRCNN framework is proposed to improve the generalisation ability on different pedestrian detection datasets, which can match the detection quality of some custom architectures. However, a similar version of pedestrian detection needs to be developed. The PRNet and the APGAN frameworks are the two latest frameworks that can improve generalisation ability. The PRNet framework aims to solve problems such as occluded pedestrian detection. Although the cross-domain performance of PRNet is good on some general pedestrian datasets, it benefits from the occluded features learnt and may not lead to a universal good result on other datasets or tasks. The APGAN framework aims to improve the generalisation ability of pretrained detectors on new datasets. However, the data augmentation method in APGAN needs many other hyperparameters, including the sizes, scales, numbers, and localizations of generated pedestrians, which limits the robustness of the framework. To present a general and robust domain adaptative framework for pedestrian detection, the HQA-Trans framework proposes a domain style conversion method to generate the intermediate domain from the source domain to the target domain. This intermediate domain will be used for cross-domain pedestrian detection, and the image quality of the generated images is guaranteed. Compared with other work, this method directly reduces the style difference between domains and could more effectively improve the generalization ability of different pedestrian detection datasets.

Miss rate that is marked as MR^{-2} in Tables 6 and 7 refers to the proportion of pedestrians that are not correctly detected. The lower the rate, the better the detection performance. According to the experimental comparison results in Table 6, it can be seen that the miss rate of our method is 9%–10% lower than the Adapted FasterRCNN framework, 39%–40% lower than the ACF framework, about 6%–7% lower than the PRNet framework, 12%–13% lower than the ALFNet framework, and 8%–9% lower than the APGAN framework. It can be seen from Table 7 that HQA-Trans outperforms ACF by 21%–22% and outperforms CSP by 4%–5%. The above mentioned information show that the proposed HQA-Trans framework has effectively improved the detection performance on the basis of existing work.

5 | CONCLUSION

Based on the research background of unsupervised cross-domain pedestrian detection with the introduction of the UNIT framework, this study explains that due to the instability of the generation network, the generated images in the intermediate domain will be distorted. Thus, a new framework of HQA-Trans is proposed. The proposed HQA-Trans mainly integrates an image quality assessment index into the UNIT framework, so that the generated images can perform domain style conversion retaining the contents and structure characteristics of the source images. The advantages of selecting the SSIM index are that it is a full-reference image quality assessment index and the source domain images can be directly selected as the reference images, which have a high degree of security. At the same time, the index can be adjusted appropriately to avoid being affected by the image style of the source domain, which may hinder the conversion of domain style. Based on the above, the generated domain is of high quality and high training efficiency. Furthermore, unsupervised cross-domain pedestrian detection is performed on some benchmark datasets. The experimental results show that the proposed HQA-Trans framework can effectively improve unsupervised cross-domain pedestrian detection performance. Compared with other related works, it can be found that the proposed HQA-Trans framework can achieve much better detection results. Finally, we think that there are two further prospects in the future: (1) Fuse the HQA-Trans framework into other generation frameworks and (2) Try some other image quality assessment indexes, which may lead to faster and better performance.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (62006239) and the joint Funds of National Natural Science Foundation of China (U19A2083).

CONFLICT OF INTEREST

There are no conflicts of interest to declare.

PERMISSION TO REPRODUCE MATERIALS FROM OTHER SOURCES

None.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at https://urldefense.com/v3/__http://www.vision.caltech.edu/

Image_Datasets/CaltechPedestrians/__;!!N11eV2iwtfs!5kZegyNUil4QE_lbD2kl1R_4K0qKGoxyRdlvuwfe8G8MF1aYROIBqetKopV-$ and https://urldefense.com/v3/__https://www.cityscapes-dataset.com/__;!!N11eV2iwtfs!5kZegyNUil4QE_lbD2kl1R_4K0qKGoxyRdlvuwfe8G8MF1aYROIBqbpsPBEG$.

ORCID

Gelin Shen https://orcid.org/0000-0001-9293-7682

REFERENCES

1. Liu, S., et al.: A novel data augmentation scheme for pedestrian detection with attribute preserving gan. Neurocomputing. 401, 123–132 (2020)
2. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. Adv. Neural Inf. Process. Syst. 30, 700–708 (2017)
3. Lee, H.Y., et al.: Diverse image-to-image translation via disentangled representations. Int. J. Comput. Vis. 128(10), 2402–2417 (2020)
4. Bhattacharjee, D., et al.: Dunit: detection-based unsupervised image-to-image translation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4787–4796 (2020)
5. Hsu, H.K., et al.: Progressive domain adaptation for object detection. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 749–757 (2020)
6. Wang, Z., et al.: Image quality assessment: from error measurement to structural similarity (2004)
7. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017)
8. Huang, X., et al.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 179–196 (2018)
9. Mo, S., Cho, M., Shin, J.: Instagan: instance-aware image-to-image translation. In: International Conference on Learning Representations (2018)
10. Dai, J., et al.: R-fcn: object detection via region-based fully convolutional networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 29, pp. 379–387 (2016)
11. Redmon, J., et al.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
12. Duan, K., et al.: Centernet: keypoint triplets for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), p. 6568–6577 (2019)
13. Liu, W., et al.: High-level semantic feature detection: a new perspective for pedestrian detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5187–5196 (2019)
14. Adel, T., Zhao, H., Wong, A.: Unsupervised domain adaptation with a relaxed covariate shift assumption. In: AAAI, pp. 1691–1697 (2017)
15. Chen, Y., et al.: Domain adaptive faster R-CNN for object detection in the wild. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3339–3348 (2018)
16. RoyChowdhury, A., et al.: Automatic adaptation of object detectors to new domains using self-training. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 780–790 (2019)
17. PerezOrtiz, M., et al.: From pairwise comparisons and rating to a unified quality scale. IEEE Trans. Image Process. 29, 1139–1151 (2019)
18. Ma, K., et al.: Waterloo exploration database: new challenges for image quality assessment models. IEEE Trans. Image Process. 26(2), 1004–1016 (2017)
19. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. IEEE Signal Process. Lett. 20(3), 209–212 (2013)
20. Kim, J., Nguyen, A.D., Lee, S.: Deep cnn-based blind image quality predictor. IEEE Trans. Neural Network. 30(1), 11–24 (2019)
21. Prashnani, E., et al.: Pieapp: perceptual image-error assessment through pairwise preference. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1808–1817 (2018)
22. Zhang, W., et al.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circ. Syst. Video Technol. 30(1), 36–47 (2020)
23. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
24. Zhang, S., Benenson, R., Schiele, B.: Citypersons: a diverse dataset for pedestrian detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4457–4465 (2017)
25. Dollar, P., et al.: Pedestrian detection: a benchmark. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 304–311 (2009)
26. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016)
27. Dollar, P., et al.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. 36(8), 1532–1545 (2014)
28. Song, X., et al.: Progressive refinement network for occluded pedestrian detection. In: European Conference on Computer Vision, pp. 32–48 (2020)
29. Liu, W., et al.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 643–659 (2018)

How to cite this article: Shen, G., et al.: HQA-Trans: An end-to-end high-quality-awareness image translation framework for unsupervised cross-domain pedestrian detection. IET Comput. Vis. 16(3), 218–229 (2022). https://doi.org/10.1049/cvi2.12081