International Journal of Computer Vision (2021) 129:1121–1138
https://doi.org/10.1007/s11263-020-01412-0

A Shape Transformation-based Dataset Augmentation Framework for Pedestrian Detection

Zhe Chen[1] · Wanli Ouyang[1] · Tongliang Liu[1] · Dacheng Tao[1]

Received: 9 December 2019 / Accepted: 30 November 2020 / Published online: 9 January 2021

Abstract
Deep learning-based computer vision is usually data-hungry. Many researchers attempt to augment datasets with synthesized data to improve model robustness. However, the augmentation of popular pedestrian datasets, such as Caltech and Citypersons, can be extremely challenging because real pedestrians are commonly in low quality. Due to the factors like occlusions, blurs, and low-resolution, it is significantly difficult for existing augmentation approaches, which generally synthesize data using 3D engines or generative adversarial networks (GANs), to generate realistic-looking pedestrians. Alternatively, to access much more natural-looking pedestrians, we propose to augment pedestrian detection datasets by transforming real pedestrians from the same dataset into different shapes. Accordingly, we propose the Shape Transformation-based Dataset Augmentation (STDA) framework. The proposed framework is composed of two subsequent modules, *i.e.* the shape-guided deformation and the environment adaptation. In the first module, we introduce a shape-guided warping field to help deform the shape of a real pedestrian into a different shape. Then, in the second stage, we propose an environment-aware blending map to better adapt the deformed pedestrians into surrounding environments, obtaining more realistic-looking pedestrians and more beneficial augmentation results for pedestrian detection. Extensive empirical studies on different pedestrian detection benchmarks show that the proposed STDA framework consistently produces much better augmentation results than other pedestrian synthesis approaches using low-quality pedestrians. By augmenting the original datasets, our proposed framework also improves the baseline pedestrian detector by up to 38% on the evaluated benchmarks, achieving state-of-the-art performance.

Keywords Pedestrian detection · Dataset augmentation · Pedestrian rendering

1 Introduction

With the introduction of large-scale pedestrian datasets (Dollár et al. 2009; Dollar et al. 2012; Zhang et al. 2017; Geiger et al. 2013), deep convolutional neural networks (DCNNs) have achieved promising detection accuracy. However, the trained DCNN detectors may not be robust enough due to the issue that negative background examples greatly exceed positive foreground examples during training. Recent studies have confirmed that DCNN detectors trained with the limited foreground examples can be vulnerable to difficult objects which have unexpected states (Huang and Ramanan 2017) and diversified poses (Alcorn et al. 2018).

To improve detector robustness, besides designing new machine learning algorithms, many researchers attempted to augment training datasets by generating new foreground examples. For instance, Huang *et al.* (Huang and Ramanan 2017) used a 3D game engine to simulate pedestrians and adapt them into the pedestrian datasets. Other studies (Ma et al. 2018; Siarohin et al. 2018; Zanfir 2018; Ge et al. 2018) attempted to augment the person re-identification datasets by transferring the poses of pedestrians using generative adversarial networks (GANs). Despite progress, it is still

Communicated by Cha Zhang.

This work was supported by Australian Research Council Projects FL-170100117, IH-180100002, IC-190100031, DP200103223, and Australian Medical Research Future Fund MRFAI000085.

✉ Dacheng Tao
dacheng.tao@sydney.edu.au

Zhe Chen
zhe.chen1@sydney.edu.au

Wanli Ouyang
wanli.ouyang@sydney.edu.au

Tongliang Liu
tongliang.liu@sydney.edu.au

[1] University of Sydney, Sydney, NSW, Australia

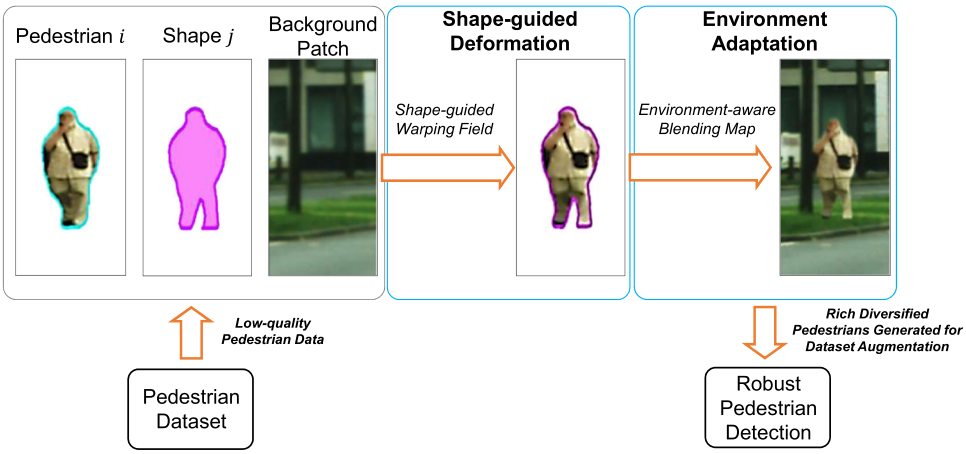


Fig. 1 We propose the shape transformation-based dataset augmentation framework for pedestrian detection. In the framework, we subsequently introduce the shape-guided warping field to deform pedestrians and the environment-aware blending map to adapt the deformed pedestrians into background environments. Our proposed framework can effectively generate more realistic-looking pedestrians for augmenting pedestrian datasets in which real pedestrians are usually in low-quality. Best view in color

very challenging to adequately apply existing augmentation approaches on the common pedestrian detection datasets. First, synthesizing pedestrians using external platforms like the 3D game engines may introduce a significant domain gap between synthesized pedestrians and real pedestrians, limiting the overall benefits for the generated pedestrians to improve the model robustness for detecting real pedestrians. Moreover, regarding the methods that utilize GANs to render pedestrians, they generally require rich appearance details from paired training images to help define the desired output of generative networks during training procedures. However, in common pedestrian detection datasets like Caltech (Dollar et al. 2012) and CityPersons (Zhang et al. 2017), pedestrians are usually in low quality due to the factors like heavy occlusions, blurry appearance, and low-resolution caused by small sizes. As a result, these available real pedestrians only provide extremely limited amount of appearance details that can be used for training generative networks. Without sufficient description of the desired appearance of synthesized pedestrians, we can show in our experiments that current GAN-based methods only generate less realistic or even corrupted pedestrians using very low-quality pedestrians from common pedestrian detection datasets.

By addressing above issues, we propose to augment pedestrian datasets by transforming real pedestrians from the same dataset according to different shapes (*i.e.* segmentation masks in this study) rather than rendering new pedestrians. Our motivation comes from the following observations. First, unlike existing methods that require sufficient appearance details to define the desired output, it is much easier to access rich pixel-level shape deformation supervision which defines the deformation from a *shape* to another shape, if only low-quality pedestrian examples are available in the datasets. The learned deformation between shapes can guide the deformation of appearances of the real pedestrians, avoiding the requirement of detailed supervision information to directly define the transformed appearances. In addition, since the shape information can naturally distinguish foreground areas from background areas, we can simply focus on adapting synthesized foreground appearances into background environments, avoiding the rick of further generating unnatural background environments together with the synthesized pedestrians as required in current GAN-based approaches. Last but not the least, we find that transforming real pedestrians based on different shapes can effectively increase foreground sample diversity while still maintaining the appearance characteristics of real pedestrians adequately.

Based on these observations, we devise a Shape Transformation based Dataset Augmentation (STDA) framework to fulfill the pedestrian dataset augmentation task more effectively. Figure 1 presents an overview of our framework. In particular, the framework first deforms a real pedestrian into a similar pedestrian but with a different shape and then adapts the shape-deformed pedestrians into surrounding environments on the image to be augmented. In the STDA framework, we introduce a shape-guided warping field, which is a set of vectors that define the warping operation between shapes, to further define an appropriate deformation between the shapes and the appearances of the real pedestrians. Moreover, we introduce an environment-aware blending map to help the shape-deformed pedestrians better blend into various background environments, delivering more realistic-looking pedestrians on the image.

In this study, our key contributions are listed as follows:

– We propose a shape transformation-based dataset augmentation framework to augment the pedestrian detection datasets and improve pedestrian detection accuracy.

To the best of our knowledge, we are the first that apply the shape-transformation-based data synthesis methodology for pedestrian detection.
- We propose the shape-guided warping field to help define a proper shape deformation procedure. We also introduce an environment-aware blending map to better adapt the shape-transformed pedestrians into different backgrounds, achieving better augmentation results on the image.
- We introduce a shape constraining operation to improve shape deformation quality. We also apply a hard positive mining loss to take advantage of the concepts of hard mining technology and further magnify the benefits of the synthesized pedestrians for improving detection robustness.
- Our proposed framework is promising for generating pedestrians, especially when using low-quality examples. Comprehensive evaluations on the famous Caltech (Dollar et al. 2012) and CityPersons (Zhang et al. 2017) benchmarks validate that our proposed framework can generate more realistic-looking pedestrians than existing methods using low-quality data. With pedestrian datasets augmented by our framework, we promisingly boost the performance of the baseline pedestrian detector, accessing superior performance to other cutting-edge pedestrian detectors.

2 Related Work

2.1 Pedestrian Detection

Pedestrian is critical in many applications such as robotics and autonomous driving (Enzweiler and Gavrila 2008; Dollár et al. 2009; Dollar et al. 2012; Zhang et al. 2016c) and downstream tasks like tracking, scene segmentation, and key point estimation (Chen et al. 2017, 2019; Zhang et al. 2020). Traditional pedestrian detectors generally use hand-crafted features (Viola et al. 2005; Ran et al. 2007) and adopt human part-based detection strategy (Felzenszwalb et al. 2010b) or cascaded structures (Felzenszwalb et al. 2010a; Bar-Hillel et al. 2010; Felzenszwalb et al. 2008). Recently, by taking advantages of large-scale pedestrian datasets (Dollár et al. 2009; Dollar et al. 2012; Zhang et al. 2017; Geiger et al. 2013; Loy et al. 2019), researchers have greatly improved the pedestrian detection performance with DCNNs (Simonyan and Zisserman 2014; He et al. 2016; Ouyang and Wang 2013; Ouyang et al. 2017). Among the DCNN detectors, two-stage detection pipelines (Ouyang and Wang 2013; Ren et al. 2015; Li et al. 2018; Cai et al. 2016; Zhang et al. 2016b; Du et al. 2017) usually perform better than single-stage detection pipelines (Liu et al. 2016; Redmon et al. 2016; Lin et al. 2018). Despite progress, the issue that foreground and background examples are extremely unbalanced in pedestrian datasets still affects the robustness of the DCNN detectors adversely. Current pedestrian detectors could still be fragile to even small transformation of pedestrians. To tackle this problem, many researchers tend to augment the datasets by synthesizing new foreground data.

2.2 Simulation-based Dataset Augmentation

To achieve dataset augmentation, researchers have used 3D simulation platforms to synthesize new examples for the datasets. For example, (Lerer et al. 2016; Ros et al. 2016) used a 3D game engine to help build new datasets. More related studies used the 3D simulation platforms to augment pedestrian-related datasets. In particular, (Pishchulin et al. 2011; Hattori et al. 2015) employed a game engine to synthesize training data for pedestrian detection. In addition, (Huang and Ramanan 2017) applied a GAN to narrow the domain gap between the 3D simulated pedestrians and the natural pedestrians to augment pedestrian datasets, but this method brings limited improvement on common pedestrian detection, suggesting that the domain gap is still large. However, there is still a significant domain gap between simulated pedestrians and real pedestrians. Such gap could further pose negative effects on DCNN detectors, making the augmented datasets deliver incremental improvements on pedestrian detection.

2.3 GAN-based Dataset Augmentation

Recently, with several improvements (Radford et al. 2015; Arjovsky et al. 2017; Gulrajani et al. 2017), GANs (Goodfellow et al. 2014) have shown great benefits on synthesis-based applications such as image-to-image translation (Isola et al. 2017; Liu et al. 2017a; Isola et al. 2017; Zhu et al. 2017) and skeleton-to-image generation (Villegas et al. 2017; Yan et al. 2017).

In the literature of person re-identification task, many works attempted to transfer the poses of real pedestrians to deliver diversified pedestrians for the augmentation. For instance, (Liu et al. 2018; Ma et al. 2018; Siarohin et al. 2018; Zanfir 2018; Ge et al. 2018; Zhang et al. 2017; Ma et al. 2017) introduced various techniques to transform the human appearance according to 2D or 3D poses and improve the person re-identification performance. (Vobecky et al. 2019) proposed a novel approach to generate pedestrians according to different poses. The synthesis results are promising and rare pedestrian situations can be simulated. In practice, these methods require accurate and reliable pose information or paired training images that contain rich appearance details to achieve successful transformation. However, existing widely used pedestrian datasets like Caltech provide neither pose annotations nor paired appearance information for training

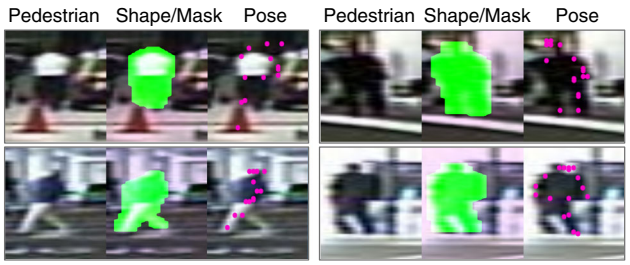


Fig. 2 Some examples showing that the shape estimation results are more accurate than pose estimation results on low-quality images using the same Mask RCNN model. Best view in color

GANs. Furthermore, in current pedestrian datasets, a large number of small pedestrians whose appearances are usually in low quality can make existing pose estimators difficult to deliver reasonable predictions. Figure 2 shows some examples describing that the poses of low-quality pedestrians are much more unstable than the masks estimated using the same Mask RCNN (He et al. 2017) detector. As a result, it is quite infeasible to seamlessly apply these pose transfer models for augmenting current pedestrian datasets.

In pedestrian detection, some studies have introduced specifically designed GANs for the augmentation. As an example, (Ouyang et al. 2018b) modified the pix2pixGAN (Isola et al. 2017) to make it more suitable for the pedestrian generation, but this method lacks a particular mechanism that helps produce diversified pedestrians and the method still delivers poor generation results based on low-quality data. In the study (Lee et al. 2018), authors introduced an end-to-end trainable neural network to fulfill the task for placing new pedestrian masks and vehicle masks in an urban scene, but it does not generate transformed pedestrian appearances to augment datasets. Also, (Liu et al. 2019) developed an effective unrolling mechanism that jointly optimizes a generative model and a detector to improve detection performance by generating new data to datasets with limited training examples. This approach directly generates pedestrian appearances from noise, while our method mainly transforms the shapes of real pedestrians to achieve better augmentation performance on low-quality data.

In this study, we propose that transforming pedestrians from the original dataset by altering their shapes can produce diversified and much more lifelike pedestrians without requiring rich appearance details for supervision.

3 Shape Transformation-based Dataset Augmentation Framework

3.1 Problem Definition

Data augmentation technique, commonly formulated as transformations of raw data, has been used to access the vast majority of the state-of-the-art results in image recognition. The data augmentation is intuitively explained as to increase the training data size and as a regularizer that can model hypothesis complexity (Goodfellow et al. 2016; Zhang et al. 2016a; Dao et al. 2019). In particular, the hypothesis complexity can be used to measure the generalization error, which is the difference between the training and test errors, of learning algorithms (Vapnik 2013; Liu et al. 2017b). Larger hypothesis complexity usually implies a larger generalization error and vice versa. In practice, a small training error and a small generalization error is favoured to guarantee a small test error. As a result, the data augmentation is especially useful for deep learning models which are powerful in maintaining a small training error but has a large hypothesis complexity. It has been empirically demonstrated that data augmentation operations can greatly improve the generalization ability of deep models (Cireşan et al. 2010; Dosovitskiy et al. 2015; Sajjadi et al. 2016).

In this study, the overall goal is to devise a more effective dataset augmentation framework to improve pedestrian detection models. The framework is supposed to generate diversified and more realistic-looking pedestrian examples to enrich the corresponding datasets in which real pedestrians are usually in very low-quality. We achieve this goal by transforming real pedestrians into different shapes rather than rendering new pedestrians. Firstly, using a deformation operation, we properly transform the shapes of pedestrians into various shapes to enrich the pedestrian appearances. The deformation introduces appropriate noises to help regularize deep models rather than existing methods like PS-GAN (Ouyang et al. 2018b) that may distract deep models by producing less realistic training examples. Secondly, we apply adequate environment adaptation to better blend the generated pedestrians into different background areas. This minimizes the risk of producing obvious unnatural artifacts that could affect performance while keeping the rich diversity of generated pedestrian appearances. Therefore, our method can be effective for regularizing the hypothesis complexity. This is empirically justified by our experiments which show that using our method to augment datasets can significantly improve pedestrian detection performance of the baseline model and out-perform other augmentation methods.

Formally, suppose $\mathbf{z}_i$ is an image patch containing a real pedestrian in the dataset and $\mathbf{s}_i$ is its extracted shape or segmentation mask. Here, we refer the shape or "mask" $\mathbf{s}_i$ of a pedestrian $\mathbf{z}_i$ as a set of labels, denoted as $s_i(x, y)$ that distinguish foreground areas from background areas within the pedestrian patch, where (x, y) represent coordinates on the image: $s_i(x, y) = 1$ for the location (x, y) being on the foreground and $s_i(x, y) = 0$ for the location (x, y) being on the background. Denote $\mathbf{s}_j$ as a different shape which can be obtained based on another real pedestrian's shape. In this study, we implement a shape transformation-based dataset augmentation function, denoted as f_{STDA}, to gener-

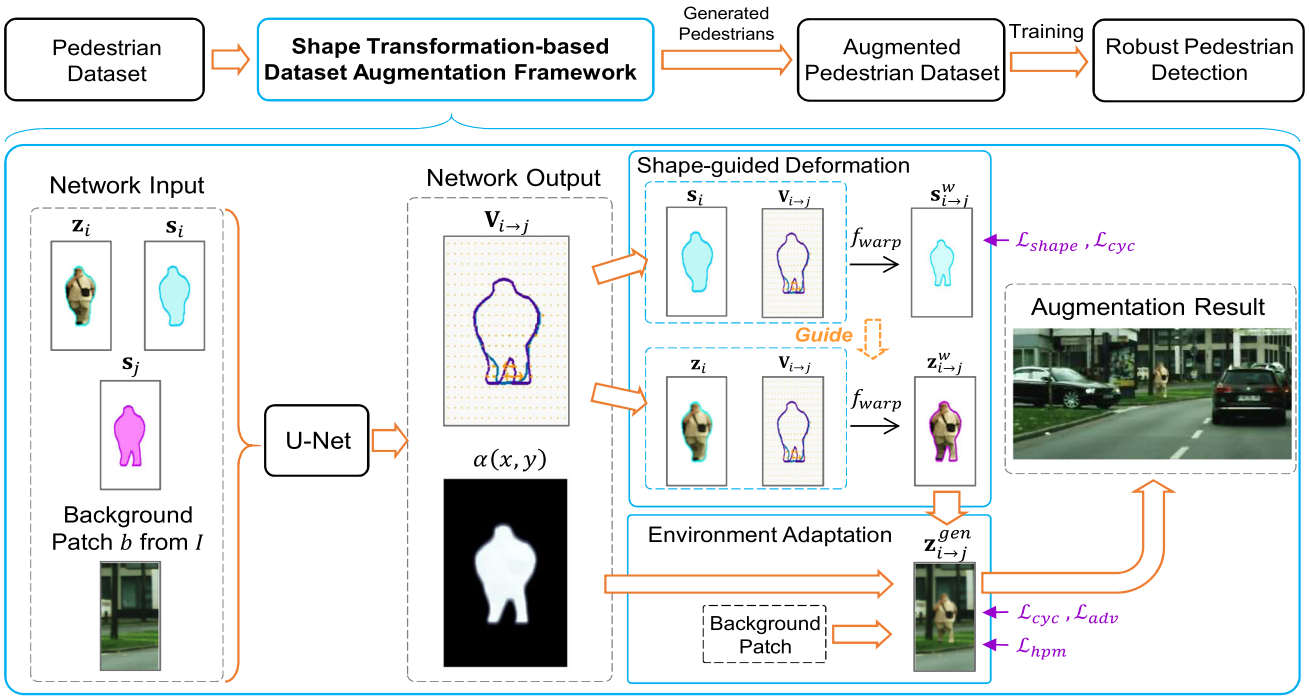


Fig. 3 Overview of the proposed shape transformation-based dataset augmentation framework for pedestrian datasets with low-quality pedestrian data. In particular, we introduce the shape-guided warping field, $\mathbf{V}_{i\to j}$, and the environment-aware blending map, (x, y), to respectively help implement the shape-guided deformation and the environment adaptation, obtaining the deformed shape $\mathbf{s}^{w}_{i\to j}$, the deformed pedestrian $\mathbf{z}^{w}_{i\to j}$, and the transformation result $\mathbf{z}^{gen}_{i\to j}$. By placing the $\mathbf{z}^{gen}_{i\to j}$ into the I, we can effectively augment the original image. In practice, we employ a U-Net to predict both of the $\mathbf{V}_{i\to j}$ and $\alpha(x, y)$. Training losses for the U-Net include $\mathcal{L}_{shape}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{cyc}$, and $\mathcal{L}_{hpm}$. Best view in color

ate a new pedestrian by transforming a real pedestrian into a new pedestrian with a more realistic-looking appearance but with another shape $\mathbf{s}_j$ for the augmentation:

$$\mathbf{z}^{gen}_{i\to j} = f_{STDA}(\mathbf{z}_i, \mathbf{s}_i, \mathbf{s}_j, I), \tag{1}$$

where $\mathbf{z}^{gen}_{i\to j}$ is a patch containing the newly generated pedestrian $\mathbf{z}_i$ by transforming its shape $\mathbf{s}_i$ into $\mathbf{s}_j$, and I is the image to be augmented.

3.2 Framework Overview

In pedestrian detection datasets, it is difficult to access sufficient appearance details to define the desired $\mathbf{z}^{gen}_{i\to j}$, making it extremely challenging to generate realistic-looking pedestrians using low-quality appearance. To properly implement the f_{STDA}, we decompose the pedestrian generation task into two sub-tasks, *i.e.* shape-guided deformation and environment adaptation. The first task focuses on varying the appearances to enrich data diversity, and the second task mainly adapts and blends the deformed pedestrians into different environments. More specifically, we first deform the pedestrian image $\mathbf{z}_i$ into a new one with similar appearance but a different shape $\mathbf{s}_j$. We define the deformation according to the transformation from $\mathbf{s}_i$ into $\mathbf{s}_j$. Then, we adapt the deformed pedestrian image into some background environments on the image I. Denote by f_{SD} the function that implements the shape-guided deformation, and denote by f_{EA} the function that implements the environment adaptation. The proposed framework implements f_{STDA} as follows:

$$f_{STDA}(\mathbf{z}_i, \mathbf{s}_i, \mathbf{s}_j, I) = f_{EA}(f_{SD}(\mathbf{z}_i, \mathbf{s}_i, \mathbf{s}_j), I). \tag{2}$$

Figure 3 shows a detailed architecture of the proposed framework. As illustrated in the figure, we introduce a shape-guided warping field, denoted as $\mathbf{V}_{i\to j}$, to help implement the shape-guided deformation function. The warping field is formulated as the assignment of vectors on the image plane for warping between shapes. With the help of $\mathbf{V}_{i\to j}$, the deformation between different shapes can guide the deformation of appearances of real pedestrians. We also propose to apply the environment-aware blending map to achieve environment adaptation. We define the blending map as a set of weighting parameters to fuse foreground pixel values with background pixel values. We use $\alpha(x, y)$ to represent an entry of the blending map located at position (x, y). After better adapting the shape-deformed pedestrian into the background environments, we obtain diversified and more realistic-looking pedestrians to augment pedestrian detection datasets. In prac-

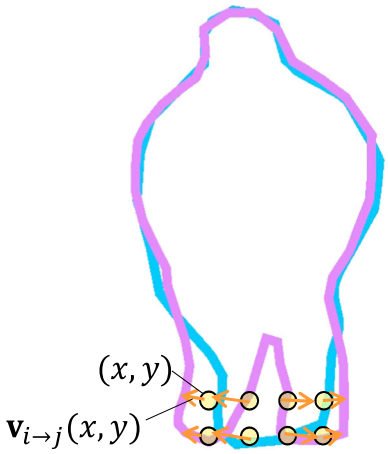


Fig. 4 A detailed example of the shape-guided warping field $\mathbf{V}_{i\to j}$ that deforms the shape $\mathbf{s}_i$ (colored in blue) into the shape $\mathbf{s}_j$ (colored in purple). (x, y) represent the 2D coordinates of a pixel on the image plane. $\mathbf{v}_{i\to j}(x, y)$ represent a vector that describes the 2D deformation offsets. Best view in color

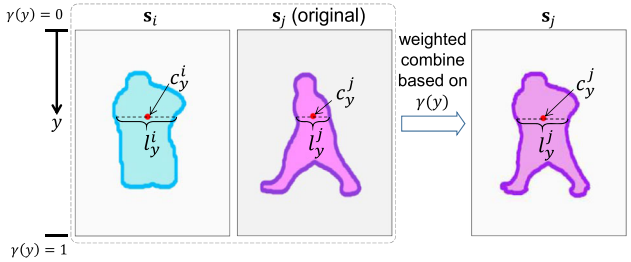


Fig. 5 An example of the shape constraining operation. In particular, we combine the shape $\mathbf{s}_j$ with $\mathbf{s}_i$ based on a weighting function $\gamma(y)$. c_y^j and l_y^j respectively denote the middle point and the width of the foreground areas on the line whose vertical offset is y and on the shape $\mathbf{s}_j$. c_y^i and l_y^i are the corresponding middle point and the width of foreground areas on the shape $\mathbf{s}_i$. The $\gamma(y)$ is a linear function of y, controlling the combination of $\mathbf{s}_i$ and $\mathbf{s}_j$. Best view in color

tice, we can employ a single end-to-end U-Net (Ronneberger et al. 2015) to help fulfill the both sub-tasks in a single pass. The employed network takes as input the pedestrian patch $\mathbf{z}_i$, its shape $\mathbf{s}_i$, the target shape $\mathbf{s}_j$, and a background patch from I, and then predicts both of the $\mathbf{V}_{i\to j}$ and the $\alpha(x, y)$. Although it is more intuitive to learn the shape-guided warping field and the environment-aware blending map separately, we simply find in practice that the U-Net has the ability to learn the knowledge for both tasks jointly. The effects of learning jointly and separately are the same. Learning jointly greatly simplifies the processing framework and saves computational resources and required parameters. Therefore, we choose to fuse the learning of both functions by feeding all the necessary input information to the U-Net at the same time.

3.2.1 Shape-guided Deformation

In this study, we implement deformation according to warping operations. In order to obtain a detailed description about warping operations, we introduce the *shape-guided warping field* to further help deform pedestrians. Denote by $\mathbf{v}_{i\to j}(x, y)$ the warping vector located at the (x, y) that helps warp the shape $\mathbf{s}_i$ into the shape $\mathbf{s}_j$. The set of these warping vectors, *i.e.* $\mathbf{V}_{i\to j} = \{\mathbf{v}_{i\to j}(x, y)\}$, then forms a shape-guided warping field. An example of this warping field can be found in Fig. 4, where the warping field helps deform the $\mathbf{s}_i$ (colored in blue) into the $\mathbf{s}_j$ (colored in purple). Then, suppose f_{warp} is the function that warps the input image patch according to the predicted warping field, we then implement the f_{SD} by:

$$\mathbf{z}_{i\to j}^{w} = f_{SD}(\mathbf{z}_i, \mathbf{s}_i, \mathbf{s}_j) = f_{warp}(\mathbf{z}_i; \mathbf{V}_{i\to j}), \tag{3}$$

where $\mathbf{z}_{i\to j}^{w}$ is the warped pedestrian $\mathbf{z}_i$ according to the shape $\mathbf{s}_j$. In practice, we define that each warping vector $\mathbf{v}_{i\to j}(x, y)$ is a 2D vector which contains the horizontal and vertical displacements between the mapped warping point and the original point located at (x, y). Thus, we can make the employed network directly predict the $\mathbf{V}_{i\to j}$. In addition, we implement the f_{warp} with the help of bilinear interpolation, since the bilinear interpolation can properly back-propagate gradients from $\mathbf{z}_{i\to j}^{w}$ to $\mathbf{V}_{i\to j}$, aiding the training effectively. For more details about using bilinear interpolation for warping and training, we refer readers to (Jaderberg et al. 2015; Dai et al. 2017).

To make the shape-guided warping field adequately describe the deformation between shapes, we define that the estimated warping field should warp the shape $\mathbf{s}_i$ into the shape $\mathbf{s}_j$. Suppose $\mathbf{s}_{i\to j}^{w}$ is the warped shape $\mathbf{s}_i$ according to $\mathbf{V}_{i\to j}$: $\mathbf{s}_{i\to j}^{w} = f_{warp}(\mathbf{s}_i; \mathbf{V}_{i\to j})$. Then, the desired warping field $\mathbf{V}_{i\to j}$ should make $\mathbf{s}_{i\to j}^{w}$ as close to $\mathbf{s}_j$ as possible. Since $\mathbf{s}_i$ and $\mathbf{s}_j$ can be easily obtained from the pedestrian datasets, we are able to access sufficient pixel-level supervision to train the employed network. We mainly apply the L_1 distance, $||\mathbf{s}_j - \mathbf{s}_{i\to j}^{w}||_1$, to measure the distance between $\mathbf{s}_j$ and $\mathbf{s}_{i\to j}^{w}$. The L_1 distance can then be used as the training loss for the network to learn the desired warping field that can help generate shape-transformed natural pedestrians based on Eq. 3.

Shape Constraining Operation: In practice, we observe that if the target shape $\mathbf{s}_j$ varies too much *w.r.t.* $\mathbf{s}_i$, the obtained warping field may distort the input pedestrians after warping, resulting in unnatural results that could degrade the augmentation performance. To avoid this, we apply a *shape constraining operation* on the target shape.

More specifically, we define the shape constraining operation as to constrain the target shape $\mathbf{s}_j$ by combining it with the input shape $\mathbf{s}_i$ according to a weighted function. The combination is defined on the middle point and the width of the

foreground areas in each horizontal line on the $\mathbf{s}_j$. Suppose y is the vertical offset for a horizontal line on $\mathbf{s}_j$. We respectively denote c_y^j and l_y^j as the middle point and the width of the foreground areas on the line y. Similarly, we refer c_y^i and l_y^i to the middle point and width of foreground areas on the $\mathbf{s}_i$ at line y. We then define the shape constraining operation as:

$$\begin{cases} c_y^{j'} = \gamma(y)\, c_y^j + (1-\gamma(y))\, c_y^i, \\ l_y^{j'} = \gamma(y)\, l_y^j + (1-\gamma(y))\, l_y^i, \end{cases} \tag{4}$$

where $c_y^{j'}$ and $l_y^{j'}$ represent the center and with on the constrained mask, and $\gamma(y)$ is the weight function *w.r.t.* y controlling the strictness of the constraint. According to Eq. 4, the smaller weight value $\gamma(y)$ can make the target shape contribute less to the combination result and *vice versa*. We set the parameter $\gamma(y)$ to different values for different parts of a body. In particular, we define $\gamma(y)$ as a linear function which increases from 0 to 1 with y varying from the top to the bottom. Therefore, when y becomes larger, we make the γ become larger accordingly. This allows more transformations for lower parts of a pedestrian body whose vertical offsets y are large.

We formulate the shape constraining operation according to Eq. 4 because we mainly hypothesize that varying more for the lower body of a natural pedestrian is more acceptable than varying more for the upper body. In particular, we find that changing too much the upper body of a pedestrian would generally require the change of the viewpoint for that pedestrian (*e.g.* from the side view to the front view) to obtain a natural appearance. However, the warping operations do not generate new image contents to generate the pedestrian in a different viewpoint.

Figure 5 shows a visual example of the introduced shape constraining operation when constraining $\mathbf{s}_j$ according to Eq. 4. We can observe that the proposed shape constraining operation adequately constrains the $\mathbf{s}_j$ by making the output shape closer to $\mathbf{s}_i$ for the upper body of the pedestrian and closer to $\mathbf{s}_j$ for the lower body.

3.2.2 Environment Adaptation

After the shape-guided deformation, we place the deformed pedestrians into the image I to fulfill augmentation. However, directly pasting deformed pedestrians could sometimes produce significant appearance mismatch due to the issues like the discontinuities in illumination conditions and imperfect shapes predicted by the mask extractor. To refine the generated pedestrians according to the environments, we further perform environment adaptation.

To properly blend a shape-deformed pedestrian into the image I by considering surrounding environments, we introduce an environment-aware blending map to help refine the deformed pedestrians. We formulate this refinement procedure as follows:

$$\mathbf{z}_{i\to j}^{gen} = f_{EA}(\mathbf{z}_{i\to j}^{w}, I) = \{\mathbf{z}_{i\to j}^{a}(x,y)\}, \tag{5}$$

where $\mathbf{z}_{i\to j}^{a}(x,y)$ is the environment adaptation result located at (x,y):

$$\begin{aligned} \mathbf{z}_{i\to j}^{a}(x,y) = &\Big(\mathbf{s}_j(x,y)\cdot(x,y)\Big)\cdot\mathbf{z}_{i\to j}^{w}(x,y) \\ &+\Big(1-\mathbf{s}_j(x,y)\cdot(x,y)\Big)\Big)\cdot I(x,y), \end{aligned} \tag{6}$$

where (x,y) is an entry value of the environment-aware blending map located at (x,y). Therefore, this refinement procedure as described above represents that each output pixel $\mathbf{z}_{i\to j}^{a}(x,y)$ is a weighted combination of a pixel $\mathbf{z}_{i\to j}^{w}(x,y)$ from the shape deformed pedestrian patch and a pixel $I(x,y)$ from the original image. The combination weight is computed by $\mathbf{s}_j(x,y)\cdot(x,y)$ where $\mathbf{s}_j(x,y)$ is 1 for foreground areas and 0 for background areas. An example of the estimated (x,y) can be found in Fig. 3.

In practice, it is difficult to define the desired refinement result and the desired environment-aware blending map. Therefore, we can not access appropriate supervision information to train the employed network for environment adaptation. Without supervision, we apply an adversarial loss to facilitate the employed network to learn and blend the deformed pedestrians into the environments effectively. Similar to the shape-guided warping field, we make the employed network directly predict environment-aware blending map. Note that we constrain the environment-aware blending map to prevent changing the appearance of the deformed pedestrians too much. In particular, we adopt a shifted and rescaled *tanh* squashing function to make the values of $\alpha(x,y)$ lie in a range of 0.8 and 1.2.

3.3 Objectives

Since we employ a single network to predict both the shape-guided warping field and the environment blending map, we can unify the objectives for training.

First, to obtain a proper shape-guided warping field, we introduce a shape deformation loss and a cyclic reconstruction loss. The shape deformation loss ensures that the predicted warping field satisfies the constrain as described in Sect. 3.2.1. The cyclic reconstruction loss then ensures that the deformed shape and pedestrian can be deformed back to the input shape and pedestrian. Therefore, we define that the shape deformation loss function $\mathcal{L}_{shape}$ for a pair of samples $i \to j$ is as follows:

$$\mathcal{L}_{shape} = \mathbb{E}[||\mathbf{s}_j - \mathbf{s}_{i\to j}^{w}||_1], \tag{7}$$

and the cyclic loss is defined as follows:

$$\mathcal{L}_{cyc} = \mathbb{E}[||\mathbf{s}_i - \mathbf{s}^w_{j\to i}||_1 + ||\mathbf{z}_i - \mathbf{z}^w_{j\to i}||_1], \tag{8}$$

where $\mathbf{s}^w_{j\to i}$ is the deformation result of $\mathbf{s}^w_{i\to j}$ according to $\mathbf{s}_i$ and $\mathbf{z}^w_{j\to i}$ is the deformation result of $\mathbf{z}^w_{i\to j}$ using the same warping field for computing $\mathbf{s}^w_{j\to i}$. As a result, the Eq. 7 describes the L_1-based shape deformation loss, and the Eq. 8 form the cyclic reconstruction loss.

In addition, an adversarial loss, denoted as $\mathcal{L}_{adv}$, is included to make sure that the shape-guided deformation and environment adaptation can help produce more realistic-looking pedestrian patches. Similar with typical GANs, the adversarial loss is computed by introducing a discriminator D for the employed network:

$$\mathcal{L}_{adv} = \mathbb{E}[\log D(\mathbf{z})] + \mathbb{E}[\log\left(1 - D(\mathbf{z}^{gen}_{i\to j})\right)], \tag{9}$$

where $\mathbf{z}$ refers to any real pedestrian in the dataset.

Hard Positive Mining Loss: Since our final goal is to improve the detection performance, we further apply a hard positive mining loss to magnify the benefits of the transformed pedestrians on improving detection robustness. Inspired by the study of hard positive generation (Wang et al. 2017), we attempt to generate pedestrians that are not very easy to be recognized by a RCNN detector (Girshick et al. 2014). Different from the study (Wang et al. 2017) that additionally introduced an occlusion mask and the spatial transformation operations to generate hard positives, we only introduce a loss function to help the employed network learn to produce harder positives for the RCNN detector. To compute this loss, we additionally train a RCNN, denoted as R, to distinguish pedestrian patches from background patches which do not contain pedestrians inside. Suppose $\mathcal{L}_{hpm}$ is the hard positive mining loss, then we have:

$$\begin{aligned}\mathcal{L}_{hpm} = {} & \mathbb{E}[\log(1 - R(\mathbf{z}^{gen}_{i\to j}))] + \mathbb{E}[\log R(\mathbf{z})] \\ & + \mathbb{E}[\log(1 - R(\mathbf{b}))],\end{aligned} \tag{10}$$

where $\mathbf{b}$ refers to background image patches in the dataset. Although hard mining is a well-developed technology, the contribution brought by $\mathcal{L}_{hpm}$ is to facilitate the *synthesis* for pedestrian examples that are more difficult to detect but more beneficial for training, which is different from common hard mining approaches.

The major difference between the $\mathcal{L}_{hpm}$ and the $\mathcal{L}_{adv}$ is that the R distinguishes between pedestrians patches and background patches, while D in $\mathcal{L}_{adv}$ distinguishes between true pedestrian patches and the shape-transformed pedestrian patches.

Overall Loss. To sum up, the overall training objective $\mathcal{L}$ of the network employed to help implement the proposed framework can be written as follows:

$$\mathcal{L} = \omega_1\mathcal{L}_{shape} + \omega_2\mathcal{L}_{cyc} + \omega_3\mathcal{L}_{adv} + \omega_4\mathcal{L}_{hpm}, \tag{11}$$

where $\omega_1, \omega_2, \omega_3, \omega_4$ are the corresponding loss weights. In general, we borrow the setting from the implementation of pix2pixGAN[1] and set the ω_1 and ω_3 to 100 and 1, respectively. Since we find in the experiment that the network can hardly learn a proper shape-guided warping field if ω_2 is too large, we empirically set the ω_2 to a small value, *i.e.* 0.5, in this study. Similarly, we also set ω_4 to 0.5 to make the hard positive mining loss contribute less to the overall objective. In practice, the network is obtained by minimizing the overall loss $\mathcal{L}$, the discriminator D is obtained by maximizing the $\mathcal{L}_{adv}$, and the R is obtained by maximizing the $\mathcal{L}_{hpm}$.

3.4 Dataset Augmentation

When augmenting the pedestrian datasets with the proposed framework, we attempted to sample more natural locations and sizes to place the transformed pedestrians in the image. Fortunately, pedestrian datasets deliver sufficient knowledge encoded within bounding box annotations to define these geometric statistics of a natural pedestrian. For example, in the Caltech dataset (Dollár et al. 2009; Dollar et al. 2012), the aspect ratio of a pedestrian is usually around 0.41. In addition, it is also possible to describe the bottom edge y^{box} and the height h^{box} of an annotated bounding box for a pedestrian using a linear model (Park et al. 2010): $h^{box} = ky^{box} + b$, where k and b are the coefficients. In the Caltech dataset whose images are 480 by 640, the k and b are found to be around 1.15 and -194.24. For each image to be augmented, we sample several locations and sizes according to this linear model. To avoid sampling patches with inappropriate background, we tend to constrain that the sampled boxes should not be quite different from the neighboring boxes of true pedestrians. For example, we tend to sample locations around true pedestrians (within 100 pixels), and we constrain the difference between the height of a sampled patch and the height of its nearest true pedestrian to be within 20 pixels. Then, for each sampled location and size, we run the proposed framework and put the transformation result into the image. Algorithm 1 describes the detailed pipeline of applying the proposed framework to augment pedestrians dataset. Algorithm 2 describes in details how we sample a location and a size in an image, which can reduce the risk of introducing inappropriate background by sampling around true pedestrians.

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

Algorithm 1 Pedestrian Dataset Augmentation Pipeline

Require: Natural pedestrians $\{\mathbf{z}_i\}$, the corresponding shapes $\{\mathbf{s}_i\}$, and images from original dataset $\mathcal{I} = \{I\}$.
Ensure: Augmented dataset images $\mathcal{I}$;
1: **for** each image $I \in \mathcal{I}$ **do**
2: Uniformly sample a number n from the set $\{1, 2, 3, 4, 5\}$;
3: **for** $m = 1 : n$ **do**
4: Sample a location and a size according to Algorithm 2;
5: Sample a pedestrian patch $\mathbf{z}_i$, a shape $\mathbf{s}_j$ from a different pedestrian, and a background patch cropped from the sampled location and size on I ;
6: Perform shape-guided deformation according to Eq. 3 and then perform environment adaptation according to Eq. 5, obtaining $\mathbf{z}^{gen}_{i\to j}$;
7: Place the $\mathbf{z}^{gen}_{i\to j}$ into the image I according to the sampled location and size;
8: **end for**
9: **end for**
10: **return** $\mathcal{I}$

Algorithm 2 Pedestrian Sampling Strategy

Require: Locations and sizes of ground-truth pedestrians for a frame, and the estimated k and b for computing pedestrian heights.
Ensure: Newly sampled location (x', y') and size (w', h') for the same frame;
1: **while** True **do**
2: **if** ground-truth pedestrians exist in current frame **then**
3: randomly sample a ground-truth pedestrian in the frame, and obtain its location (x, y) and size (w, h);
4: randomly sample an offset (dx, dy) *w.r.t.* the (x, y) according to a uniform distribution, such that $\sqrt{dx^2 + dy^2} \leq 100$;
5: compute the newly sampled location $(x', y') = (x, y) + (dx, dy)$;
6: compute the height according to $h' = ky' + b$, and clip its value such that $|h' - h| \leq 20$;
7: **else**
8: randomly sample a location (x', y') in the image;
9: compute the height according to $h' = ky' + b$;
10: **end if**
11: compute the width according to $w' = \alpha h'$ (*e.g.* in Caltech, $\alpha = 0.41 + \mathcal{U}(-0.1, +0.1)$, where $\mathcal{U}$ is a uniform distribution);
12: **if** h' is too small (*e.g.* less than 24 pixels) or (x', y', w', h') heavily overlaps with existing pedestrians (*e.g.* IoU > 0.3) **then**
13: continue;
14: **else**
15: break;
16: **end if**
17: **end while**
18: **return** (x', y') and (w', h')

4 Experiments

We perform comprehensive evaluation for the proposed STDA framework to augment pedestrian datasets. We use the popular Caltech (Dollár et al. 2009; Dollar et al. 2012) and CityPersons (Zhang et al. 2017) benchmarks for the evaluation.

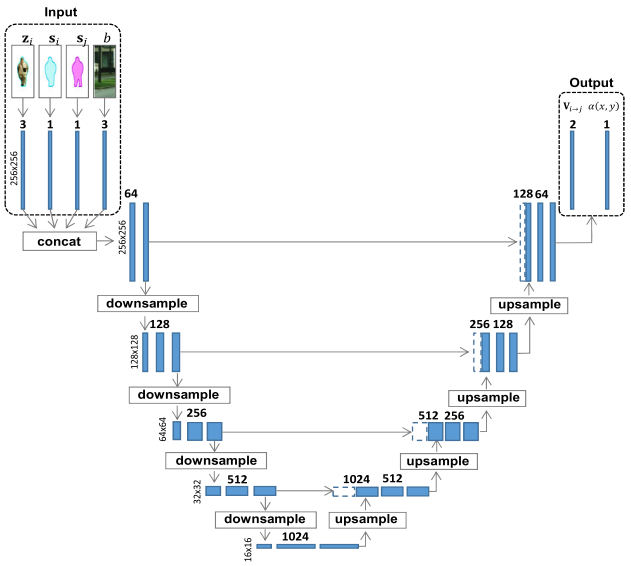


Fig. 6 Detailed structure of the employed U-net. Best view in color. Each blue rectangle represents a convolution. For each rectangle, numbers on the side represent resolution and number on the top represents channel number

In this section, we will first present the overall dataset augmentation results on evaluated datasets. Then, we will validate the improvements in improving detection accuracy of applying our proposed STDA framework to augment different datasets, comparing to other cutting-edge pedestrian detectors. Subsequently, we perform detailed ablation studies on the STDA framework to analyze the effects of different components in STDA on generating more realistic-looking pedestrians and on improving detection accuracy.

4.1 Settings and Implementation Details

For evaluation, we consider the log-average miss rates (MR) against different false positive rates as the major metric to represent pedestrian detection performance. In the Caltech, we follow the protocol of (Zhang et al. 2016b) and use around 42k images for training and 4024 images for testing. In the Citypersons, as suggested in the original study, we use 2975 images for training and perform the evaluation on the 500 images from the validation set. We apply a Mask RCNN to extract shapes on Caltech and use the annotated pedestrian masks on Citypersons. To augment the datasets, for each frame, we transform n pedestrians using our framework and n is uniformly sampled from $\{1, 2, 3, 4, 5\}$. Thus, each image has the number of positive pedestrians increased by $1 \sim 5$.

For the network employed to implement the framework, we use the U-net architecture with 8 blocks. All the input and output patches have a size of 256×256. Figure 6 shows the detailed structure of the employed U-net. Then, both the D as introduced in Eq. 9 and the R as introduced in Eq. 10 are

CNNs with 3 convolutional blocks. During optimization, we reduce the updating frequency of D and R to stabilize the training, *i.e.* we update D and R once at every 40-th update of the U-net. Learning rate is set to $1e-5$ and we perform training with 80 epochs for a dataset.

We adopt a ResNet50-based FPN detector (Lin et al. 2017) as our major baseline detector. When training this detector, we modified some default parameters according to the pedestrian detection task. First, for the region proposal network in FPN, we follow the (Zhang et al. 2016b) and only use the anchor with the aspect ratio of 2.44. We discard the 512x512 anchors in FPN because they do not contribute much to the performance. In addition, we set the batch size as 512 for both the Region Proposal Network (RPN) and the Regional-CNN (RCNN) in FPN. To reduce of false positive rates of FPN, we further set the foreground thresholds of RPN and RCNN to 0.5 and 0.7, respectively. During training, we make the length of the shorter size of input images as 720 for Caltech and as 1024 for CityPersons. Both FPN baseline and FPN trained with our methods are pre-trained on MS COCO (Lin et al. 2014) dataset to gain proper prior knowledge about people. We train the FPN detector on the Caltech with 3 epochs and on the CityPersons with 6 epochs. In general, the final performance of the baseline detector is 10.4% mean miss rate on Caltech test set and 13.9% mean miss rate on CityerPersons validation set. Note that we weight the loss values for synthesized pedestrians by a factor of 0.1, reducing the potential biases towards generated pedestrians rather than real pedestrians.

For the hyper parameters introduced in this study, like the loss weights of cyclic loss and hard positive mining loss, we mainly select them by performing grid search according to the quality of generated pedestrians and the performance of improved detectors.

In addition to FPN, we further adopt a MS CNN (Cai et al. 2016) as another baseline to evaluate our method more comprehensively. We use the released source codes to implement the MS CNN and use similar weight loss for synthesized pedestrians to train MS CNN with the proposed STDA.

4.2 Dataset Augmentation Results

We first present the pedestrian synthesis results of applying our STDA framework to augment pedestrian datasets.

4.2.1 Pedestrian Synthesis Results

In Fig. 7, we illustrate the dataset augmentation results on both the evaluated Caltech dataset and citypersons dataset. Even if some of the pedestrians are blurry and lack rich appearance details, we can observe that the shape transformed pedestrians can still be naturally blended into the environments of the image, obtaining very realistic-looking pedestrians for dataset augmentation. Furthermore, the STDA can also produce pedestrians in uncommon walking areas, such as in the middle areas of the street. This can increase the irregular foreground examples for pedestrian detection, and the model can be more robust in detecting pedestrians after augmentation. Moreover, with a similar geometry arrangement with real pedestrians, the illustrated results can demonstrate that the proposed STDA framework is effective in generating pedestrians in a similar domain with real pedestrians. Besides, our method can produce occlusion cases, *e.g.* by overlapping the generated pedestrians over real pedestrians, which can promisingly increase the amount of occlusion cases for training and thus improve the detection robustness for occlusions.

In addition, we also compare our method with another recently published powerful GAN-based data rendering technique, *i.e.* PS-GAN (Ouyang et al. 2018b), using the same background patches. We implement the PS-GAN with the codes released by its authors and follow the original training scripts to train the model. However, the original PS-GAN does not include training datasets. For fair comparison, we modified its training scripts to include the same training data as used in our method. As mentioned in the paper, there are plenty of very low-quality pedestrians in our training data. Furthermore, since we discarded irregular shapes according to predicted confidence scores of the Mask RCNN, the number of obtained pedestrians for training is relatively small. Training schemes for both our study and the PS-GAN are kept the same. Figure 8a shows some pedestrian synthesis results using existing GANs. We can find that the compared GAN-based method produces very blurry pedestrians. Besides, the generated backgrounds can be also unnatural and distorted. There could be a few reasons why the compared PS-GAN works badly. First, since PS-GAN generates pedestrians without conditioning on the quality of training examples, the mix of high-quality data and very low-quality data as used by us could confuse the PS-GAN during training and affect the quality of generated pedestrians. Also, since the number of pedestrians used here for training is relatively small, it is difficult to train the PS-GAN very thoroughly. On the contrary, as shown in Fig. 8b, our proposed STDA framework can effectively generate much more realistic and natural-looking pedestrians in different background patches. Our method achieves significantly lower score than PS-GAN, meaning that the STDA-generated pedestrians are much more similar to the true data. This illustrates its superiority over the GAN-based data rendering methods.

4.2.2 Improvements for Pedestrian Detection

Caltech: To evaluate the augmentation results of our proposed STDA framework, we first perform the evaluation on the test set of the Caltech benchmark. We evaluate the

Fig. 7 Dataset augmentation results of the proposed STDA on images from Caltech (top 2 rows) and CityPersons (bottom 3 rows), respectively. Light green bounding boxes indicate the synthesized pedestrians. The presented image patches are cropped and zoomed to better illustrate the details. Best view in color

performance gains with respect to the baseline detector to demonstrate the effectiveness. Figure 9 shows the detailed performance of our method comparing to other cutting-edge methods (Ouyang et al. 2018a; Zhang et al. 2018b; Cai et al. 2016; Li et al. 2018; Zhang et al. 2016b, 2017; Du et al. 2017; Brazil et al. 2017; Lin et al. 2018). In particular, our framework improves around 30% miss rate over the baseline FPN. By further applying the multi-scale testing, we can achieve 38% improvement, significantly outperforming other cutting-edge pedestrian detectors. Moreover, our method also delivers 3 points' improvements over another baseline detector, MS CNN. Using the multi-scale testing for MS CNN, our method further improves 3.8 points, obtaining the lowest average miss rates 6.1%. This shows that

(a) Synthesis results of PS-GAN (Ouyang et al., 2018b).

(b) Synthesis results of STDA (ours)

Fig. 8 Pedestrian synthesis results of STDA, comparing to another cutting-edge GAN-based data generation method. Synthesized pedestrians are in the middle of each image patch and background patches are kept the same. Best view in color

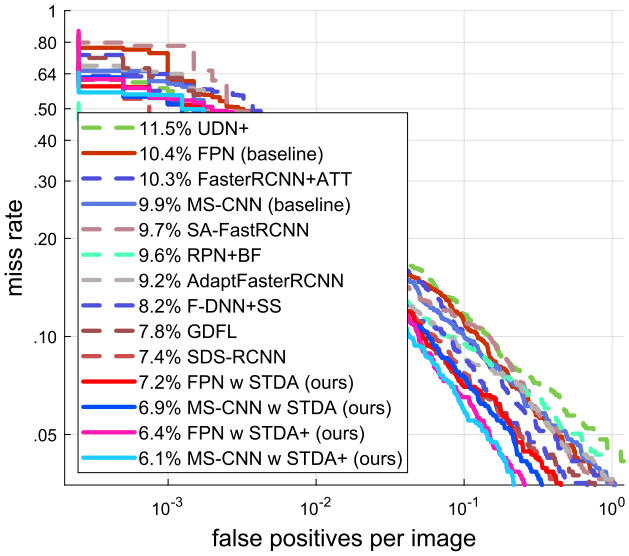


Fig. 9 Effects of the proposed STDA framework for augmenting the Caltech pedestrian dataset, comparing to other cutting-edge pedestrian detectors. "+" means multi-scale testing

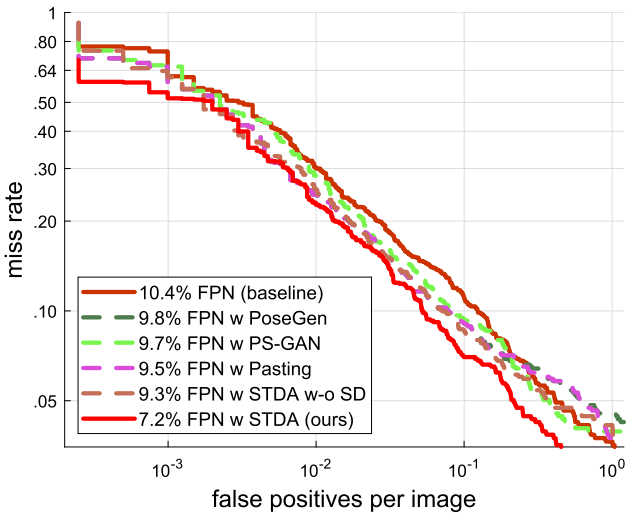


Fig. 10 Effects of the proposed STDA framework for augmenting the Caltech pedestrian dataset, comparing to other pedestrian synthesis methods such as random pasting, PoseGen (Ma et al. 2017), and PS-GAN (Ouyang et al. 2018b). We also compare our partial method, "STDA w-o SD", that only blends pedestrians into environments *without* shape deformation for ablation study

our method can consistently improve different trained detectors.

We also present qualitative results of whether using our method on the FPN detector in Fig. 11. With limited training examples in existing datasets, we can find in the figure that the baseline FPN detector produced inaccurate results (first column), false positives (second column), or false negatives (their column). On the contrary, the FPN trained with our method can correctly detect pedestrians in the corresponding images by providing more accurate boxes and less false predictions. This further demonstrates that our method is effective for improving detection performance by including more diversified training examples.

In Fig. 10, we also compare our framework with some other augmentation methods, including Pasting that directly pastes real pedestrians randomly, PS-GAN (Ouyang et al. 2018b) that generates pedestrian patches based on a pix2pixGAN (Isola et al. 2017) pipeline, and our method that only blends real pedestrians without shape deformation.

Fig. 11 Comparison of qualitative results about whether applying our method for dataset augmentation. Red boxes are ground-truths. Green dotted boxes are detection results. Best view in color

Table 1 Performance on occluded (OCC) pedestrians on Caltech test set. Best results are highlighted in **bold**. "+" means multi-scale testing

Methods/MR(%)	OCC (heavy)	OCC (partial)
UDN+ (Ouyang et al. 2018a)	70.3	28.2
FasterRCNN+ATT (Zhang et al. 2018b)	45.2	22.3
SA-FastRCNN (Li et al. 2018)	64.4	24.8
RPN+BF (Zhang et al. 2016b)	74.4	24.2
AdaptFasterRCNN (Zhang et al. 2017)	57.6	26.5
F-DNN+SS (Du et al. 2017)	53.8	15.1
GDFL (Lin et al. 2018)	43.2	16.7
SDS-RCNN (Brazil et al. 2017)	58.5	14.9
FPN (baseline) (Lin et al. 2017)	59.3	22.9
MS-CNN (baseline) (Cai et al. 2016)	59.9	19.2
FPN w STDA (ours)	41.9	17.2
FPN w STDA+ (ours)	43.5	12.4
MS-CNN w STDA (ours)	**39.8**	13.6
MS-CNN w STDA+ (ours)	40.1	**11.7**

We can find that the three other compared methods can also slightly improve the baseline detector, suggesting that augmenting pedestrian datasets with synthesized pedestrians is useful for improving detection accuracy. However, due to the unnatural pedestrians synthesized based on low-quality data as presented in Fig. 8a, improvements brought by PS-GAN is very limited. Even random pasting real pedestrians can deliver a slightly better improvements using low-quality data. Moreover, we can also observe that the compared methods have higher false positives per image than the baseline detector at low miss rates, suggesting that the baseline detector may be distracted by unnatural pedestrians to some extents. Comparing to the other compared pedestrian synthesis methods, the performance gain brought by our proposed STDA is much more significant with respect to the baseline detector, confirming that our proposed framework is much more effective in augmenting pedestrian datasets using low-quality pedestrian data. Furthermore, with the more realistic-looking pedestrians synthesized by STDA, the augmented dataset can consistently improve the baseline detector at all presented false positives per image. To further validate the idea of deforming the shapes of pedestrians to augment datasets, we perform another ablation study to evaluate our method that only blends real pedestrians into environments without deforming their shapes. The results are also presented in Fig. 10. It shows that dataset augmentation without using shape deformation delivers worse performance than our complete method, even though it improves random pasting performance. This demonstrates that the shape deformation, which enhances the diversity of synthetic pedestrians, is important to improve detection performance.

Besides overall performance, we also present performance on specific detection attributes. For example, Table 1 shows the detection accuracy on pedestrians with partial or heavy occlusions. According to the statistics, we can find that the proposed STDA can effectively reduce the average miss rate of the baseline detector for both partial and heavy occluded pedestrians, achieving favorable performance comparing to other cutting-edge pedestrian detectors. This confirms that synthesizing pedestrians with occlusions using our proposed STDA framework can promisingly help improve the detection robustness and accuracy of occluded pedestrians in test set. In addition, we also evaluate the performance of applying STDA to augment the Caltech on pedestrians with different aspect ratios in the Table 2. In particular, for the detection on the pedestrians with "typical" aspect ratios, our proposed framework is able to boost the performance of the baseline detector by up to 41%. When detecting the pedestrians with "a-typical" aspect ratios, our method also promisingly improves the baseline performance, obtaining the highest average miss rate among compared pedestrian detectors. These results demonstrate that our framework can produce rich diversified and beneficial pedestrians for the augmentation. Furthermore, Table 3 shows detection performance of pedestrians at medium or far distances. It shows that our method improves greatly on pedestrians at both distances. Since "far" pedestrians usually have small sizes (*e.g.* bounding box heights are less than 80 pixels), our method shows to be beneficial for enhancing detectors' performance on small pedestrians that are originally difficult to detect.

Table 2 Performance on pedestrians with diversified aspect ratios (AR) on Caltech test set. Best results are highlighted in **bold**. "typical" means the pedestrians with normal aspect ratios; "a-typical" means the pedestrians with unusual aspect ratios; "+" means multi-scale testing

Methods/MR(%)	AR(typical)	AR(a-typical)
UDN+ (Ouyang et al. 2018a)	7.8	14.9
FasterRCNN+ATT (Zhang et al. 2018b)	6.0	19.4
SA-FastRCNN (Li et al. 2018)	5.7	15.8
RPN+BF (Zhang et al. 2016b)	6.0	14.5
AdaptFasterRCNN (Zhang et al. 2017)	5.0	16.2
F-DNN+SS (Du et al. 2017)	5.1	13.3
GDFL (Lin et al. 2018)	4.5	14.7
SDS-RCNN (Brazil et al. 2017)	4.6	11.7
FPN (baseline) (Lin et al. 2017)	6.6	15.7
MS-CNN (baseline) (Cai et al. 2016)	6.3	15.7
FPN w STDA (ours)	4.5	11.3
FPN w STDA+ (ours)	3.9	9.9
MS-CNN w STDA (ours)	4.3	11.2
MS-CNN w STDA+ (ours)	**3.4**	**9.8**

Table 3 Performance on pedestrians with different distances on Caltech test set. Best results are highlighted in **bold**. "far" means the pedestrians at longer distances; "medium" means the pedestrians at medium distances; "+" means multi-scale testing

Methods/MR(%)	AR(medium)	AR(far)
UDN+ (Ouyang et al. 2018a)	53.8	100
FasterRCNN+ATT (Zhang et al. 2018b)	40.7	90.9
SA-FastRCNN (Li et al. 2018)	51.8	100
RPN+BF (Zhang et al. 2016b)	53.9	100
AdaptfasterRCNN (Zhang et al. 2017)	48.5	99.8
F-DNN+SS (Du et al. 2017)	33.1	77.4
GDFL (Lin et al. 2018)	32.5	71.0
SDS-RCNN (Brazil et al. 2017)	50.9	100
FPN (baseline) (Lin et al. 2017)	40.3	82.4
MS-CNN (baseline) (Cai et al. 2016)	49.1	97.2
FPN w STDA (ours)	32.6	74.4
FPN w STDA+ (ours)	**31.3**	**70.9**
MS-CNN w STDA (ours)	32.8	76.9
MS-CNN w STDA+ (ours)	31.5	75.3

CityPersons: In this section, we also report the performance on the validation set of CityPersons. The experiment settings are similar to the evaluation for the Caltech dataset except that image sizes are 1024×2048 for training and testing.

Table 4 presents the detailed statics of the evaluated methods. We can find that our framework effectively augments the original dataset and improves the performance of the baseline FPN detector. Besides, our approach also improves the MS CNN promisingly. The MS CNN trained with our approach achieves the highest single model and multi-scale testing results among compared detectors. By achieving state-of-the-art performance with our proposed framework, we can validate that our proposed framework can consistently augment different pedestrian datasets with low-quality pedestrian data.

Table 4 Performance on the validation set of CityPersons. Best results are highlighted in **bold**. "+' means multi-scale testing

Method/MR%	Reasonable	Heavy	Partial	Bare
Citypersons (Zhang et al. 2017)	15.4	–	–	–
TLL (Song et al. 2018)	14.4	52.0	15.9	9.2
RepultionLoss (Wang et al. 2018)	13.2	56.9	16.8	7.6
OR-CNN (Zhang et al. 2018a)	12.8	55.7	15.3	6.7
FPN (baseline) (Lin et al. 2017)	13.9	52.9	15.4	8.5
MS CNN (baseline) (Cai et al. 2016)	13.2	51.4	13.9	8.2
FPN w STDA (ours)	11.0	44.1	11.3	6.4
FPN w STDA+ (ours)	10.2	41.9	10.5	5.8
MS CNN w STDA (ours)	10.6	43.4	10.6	6.3
MS CNN w STDA+ (ours)	**10.0**	**41.3**	**9.9**	**5.6**

4.3 Ablation Studies

In this section, we perform comprehensive component analysis of the proposed STDA framework for both the pedestrian generation and the pedestrian detection augmentation, using the low-quality pedestrians in Caltech dataset and the Caltech benchmark for training.

4.3.1 Qualitative Study

We first evaluate the qualitative effects of different components in the STDA framework for the pedestrian generation task. In particular, we start the experiments from only using the shape-guided deformation supervised by $\mathcal{L}_{shape}$ for pedestrian generation. Then, we gradually add the shape-constraining operation, cyclic reconstruction loss $\mathcal{L}_{cyc}$, adversarial loss $\mathcal{L}_{adv}$, environment-aware blending map $\mathbf{e}(x, y)$, and hard positive mining loss $\mathcal{L}_{hpm}$ to help generate pedestrians. We present the effects of different components by generating pedestrians based on low-quality real pedestrian data in the Fig. 12. According to the presented results, we can observe that the quality of the generated pedestrians is progressively improved by introducing more components, demonstrating the effectiveness of the different components in STDA framework. More specifically, the shape constraining operation can first help the deformation operation produce less distorted pedestrians. Then, by adding the cyclic loss $\mathcal{L}_{cyc}$ and adversarial loss $\mathcal{L}_{adv}$, the obtained

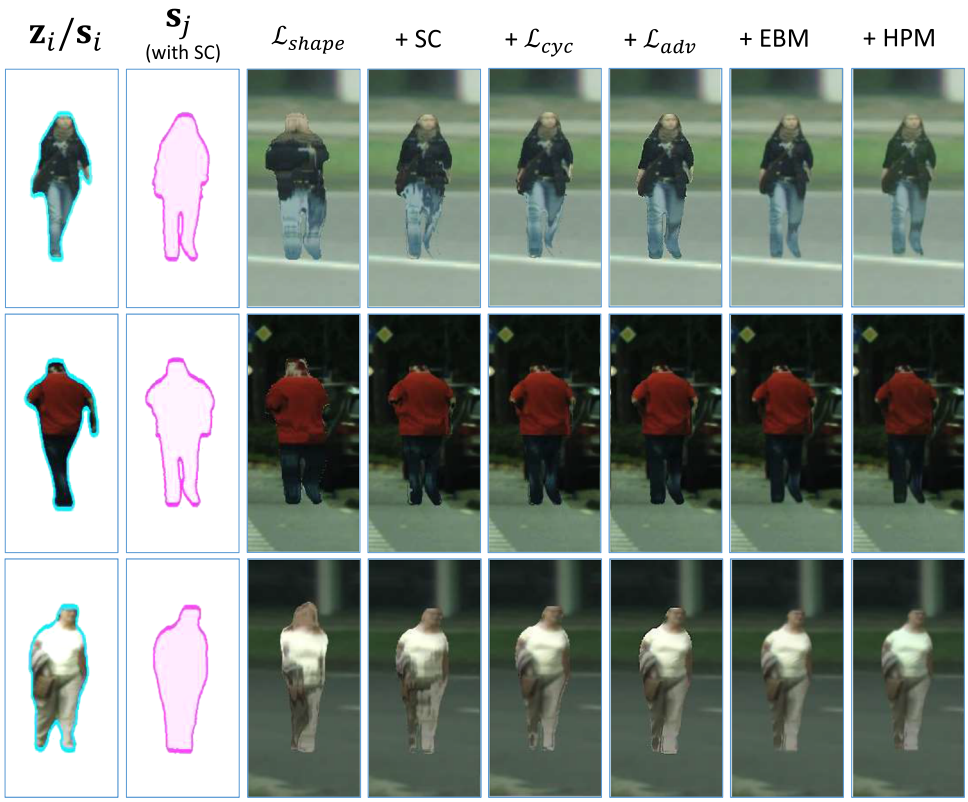
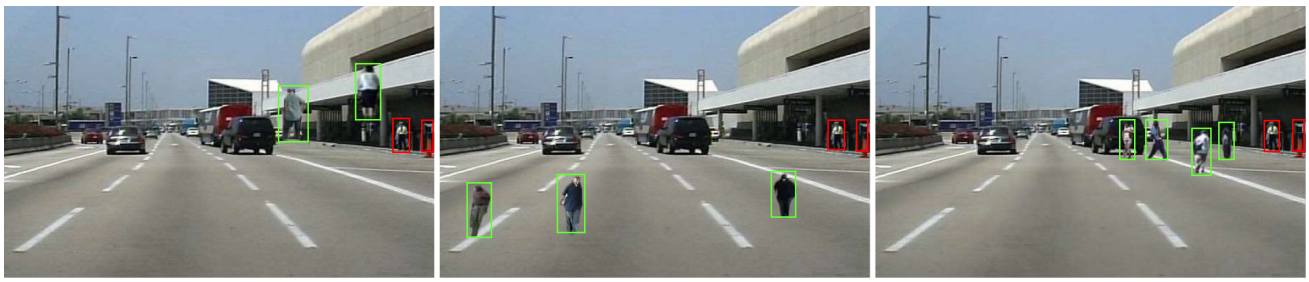


Fig. 12 Visual effects of different components in the proposed STDA framework. "SC" means shape constraining operation; "EBM" means environment-aware blending map; "HPM" means hard positive mining. Best view in color



(a) Sample pedestrians pure randomly

(b) Sample pedestrians according to the linear model

(c) Sample pedestrians according to the linear model and true pedestrians

Fig. 13 Visual effects of different sampling strategies to place synthesized pedestrians. True pedestrians are highlighted in red boxes. Sampled pedestrians are highlighted in green boxes. Frames are zoomed for better illustration. Best view in color

pedestrians become more realistic-looking in details. Subsequently, the introduced environment-aware blending map trained by $\mathcal{L}_{ebm}$ helps the transformed pedestrians better adapt into the background image patch. Lastly, the $\mathcal{L}_{hpm}$ can slightly change some appearance characteristics, such as illumination or color, to make the pedestrians less distinguishable from the environments, which actually further improve the pedestrian generation results.

In addition, we also evaluate the effects of pedestrian sampling strategy as described in Sect. 3.4 and Algorithm 2. The qualitative results are presented in Fig. 13. We compared three different schemes for sampling locations and sizes to place synthesized pedestrians: (a) we sample pedestrians in the image pure randomly; (b) we sample pedestrians in the image only according to the linear model; (c) we sample pedestrians according to the linear model and true pedestrians as described in Algorithm 2. From the presented results, we can find that scheme (a) will generate unnatural locations and sizes, making the synthesized pedestrians being placed into inappropriate background areas. Then, scheme (b) improves the effects of (a) promisingly, but the sampled sizes are still sub-optimal. The scheme (c) has the best sampling quality and can generate more appropriate locations and sizes, reducing the risk of including inappropriate background contents significantly.

Lastly, we compared different learning strategies for the proposed network, including the separate learning and the joint learning. The joint learning trains the network to pre-

Fig. 14 Visual effects of different learning strategies for training our developed network. Best view in color

Table 5 Effects of different components in the proposed STDA framework on the selected validation set on Caltech dataset. "SC" means shape constraining operation; "EBM" means environment-aware blending map; "HPM" means hard positive mining

Baseline	$\mathcal{L}_{shape}$	SC	$\mathcal{L}_{cyc}$	$\mathcal{L}_{adv}$	EBM	HPM	MR%
✓							10.77
✓	✓						10.58
✓	✓	✓					10.31
✓	✓	✓	✓				9.03
✓	✓	✓	✓	✓			8.56
✓	✓	✓	✓	✓	✓		7.73
✓	✓	✓	✓	✓	✓	✓	7.49

dict both shape-guided warping field and environment-aware blending map at the same time, while the separate learning trains two independent networks to predict the shape-guided warping field and environment-aware blending map, respectively. Figure 14 shows the synthesis results of using two learning strategies. We can find that both learning strategies produces identical synthesis performance, illustrating that the developed network for synthesizing pedestrians is insensitive to learning schemes.

4.3.2 Quantitative Study

To perform ablation studies, we split the training set of Caltech into one smaller training set and one validation set. More specifically, we collect the frames from the first four sets in the training as training images, while the frames from the last set are considered as validation images. We sample every 30-th frame in the overall dataset to set up the training/validation set. Note that this setting of training/validation set is ONLY used for ablation study.

Table 5 presents the detailed results. We can find that each of the introduced component, including shape constraining operation (SC), cyclic loss ($\mathcal{L}_{cyc}$), adversarial loss ($\mathcal{L}_{adv}$), the environment-aware blending map (EBM), and the hard positive mining (HPM), can all contribute a promising average miss rate reduction. In particular, the cyclic and adversarial loss that helps better deform pedestrians and the environment-aware blending map that helps better adapt deformed pedestrians can both greatly boost the benefits of synthesized pedestrians on improving detection accuracy. The proposed hard positive mining scheme can further improve the detection accuracy, demonstrating its effectiveness in dataset augmentation. Based on the qualitative analysis as shown in Fig. 12, we can further conclude that augmenting pedestrian datasets with more realistic-looking pedestrians can deliver better improvements on detection accuracy.

Table 6 Effects of using different sampling strategies to place synthesized pedestrians on the selected validation set on Caltech dataset. Compared strategies include: (a) Sample pedestrians pure randomly; (b) Sample pedestrians according to the linear model; (c) Sample pedestrians according to the linear model and true pedestrians

	(a)	(b)	(c)
MR%	8.07	7.68	7.49

We also studied the influence of different sampling strategies on the detection performance. Table 6 shows the results. We can find that pure random sampling that may introduce inappropriate background areas for synthesized pedestrians offers limited help for the detection performance. Introducing the linear model to sample locations and sizes for synthesized pedestrians then improves random sampling promisingly, indicating that the linear model provides a more reasonable way to place synthesized pedestrians. By further considering true pedestrians, we obtained the best performance, illustrating that using both the linear model and true pedestrians tend to avoid unnatural background areas for inserting synthesized pedestrians.

We further studied the effects of different learning strategies, *i.e.* the joint learning and the separate learning, on the selected validation set of Caltech dataset. The joint learning which is applied in our major implementation has a detection score of 7.49% log-average miss rate on the validation set. The separate learning then achieves a similar detection score which is 7.51 % log-average miss rate on the validation set. This shows that the synthesis results brought by separate learning have nearly the same benefits for improving detection compared to the joint learning.

5 Conclusions

In this study, we present a novel shape transformation-based dataset augmentation framework to improve pedestrian detection. The proposed framework can effectively deform

natural pedestrians into a different shape and can adequately adapt the deformed pedestrians into various background environments. Using low-quality pedestrian data available in the datasets, our proposed framework produces much more lifelike pedestrians than other cutting-edge data synthesis techniques. By applying the proposed framework on the two different well-known pedestrian benchmarks, *i.e.* Caltech and CityPersons, we improve the baseline pedestrian detector with a great margin, achieving state-of-the-art performance on both of the evaluated benchmarks.

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W. S., Nguyen, A. (2018). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. arXiv preprint arXiv:1811.11553

Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875

Bar-Hillel, A., Levi, D., Krupka, E., & Goldberg, C. (2010). Part-based feature synthesis for human detection. *ECCV* (pp. 127–142). New york: Springer.

Brazil, G., Yin, X., Liu, X. (2017). Illuminating pedestrians via simultaneous detection & segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. *ECCV* (pp. 354–370). New york: Springer.

Chen, Z., Li, J., Chen, Z., & You, X. (2017). Generic pixel level object tracker using bi-channel fully convolutional network. *International conference on neural information processing* (pp. 666–676). New york: Springer.

Chen, Z., Zhang, J., & Tao, D. (2019). Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, *6*(3), 693–702.

Cireşan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, *22*(12), 3207–3220.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y. (2017). Deformable convolutional networks. *In: Proceedings of the IEEE international conference on computer vision*, pp 764–773

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., Re, C. (2019). A kernel theory of modern data augmentation. *In: International Conference on Machine Learning*, pp 1528–1537

Dollár, P., Wojek, C., Schiele, B., Perona, P. (2009). Pedestrian detection: A benchmark. *In: CVPR, IEEE*, pp 304–311

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *T-PAMI*, *34*(4), 743–761.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1734–1747.

Du, X., El-Khamy, M., Lee, J., Davis, L. (2017). Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. *In: WACV, IEEE*, pp 953–961

Enzweiler, M., & Gavrila, D. M. (2008). Monocular pedestrian detection: Survey and experiments. *T-PAMI*, *12*, 2179–2195.

Felzenszwalb, P., McAllester, D., Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *In: CVPR, IEEE*, pp 1–8

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. (2010a). *Cascade object detection with deformable part models. In: CVPR, IEEE*, pp 2241–2248

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010b). Object detection with discriminatively trained part-based models. *T-PAMI*, *32*(9), 1627–1645.

Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al. (2018). Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in Neural Information Processing Systems*, *31*, 1230–1241.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237.

Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *In: Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 580–587

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. *In: NIPS*, pp 2672–2680

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C. (2017). Improved training of wasserstein gans. *In: NIPS*, pp 5767–5777

Hattori, H., Naresh Boddeti, V., Kitani, K. M., Kanade, T.(2015). Learning scene-specific pedestrian detectors without real data. *In: CVPR*, pp 3819–3827

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *In: CVPR*, pp 770–778

He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask r-cnn. *In: ICCV, IEEE*, pp 2980–2988

Huang, S., Ramanan, D. (2017). Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. *In: CVPR, IEEE*, vol 1

Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. arXiv preprint

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. *In: Advances in neural information processing systems*, pp 2017–2025

Lee, D., Liu, S., Gu, J., Liu, M. Y., Yang, M.H., Kautz, J. (2018). Context-aware synthesis and placement of object instances. *In: NeurIPS*, pp 10393–10403

Lerer, A., Gross, S., Fergus, R. (2016). Learning physical intuition of block towers by example. arXiv preprint arXiv:1603.01312

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2018). Scale-aware fast r-cnn for pedestrian detection. *TMM*, *20*(4), 985–996.

Lin, C., Lu, J., Wang, G., & Zhou, J. (2018). Graininess-aware deep feature learning for pedestrian detection. *ECCV* (pp. 732–747). New york: Springer.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. *ECCV* (pp. 740–755). New york: Springer.

Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J. (2017). Feature pyramid networks for object detection. *In: CVPR*, vol 1, p 4

Liu, J., Ni, B., Yan, Y., Zhou P., Cheng, S., Hu, J. (2018). Pose transferrable person re-identification. *In: CVPR, IEEE*, pp 4099–4108

Liu, L., Muelly, M., Deng, J., Pfister, T., Li, L. J. (2019). Generative modeling for small-data object detection. In: ICCV, pp 6073–6081

Liu, M. Y., Breuel, T., Kautz, J .(2017a). Unsupervised image-to-image translation networks. In: NIPS, pp 700–708

Liu, T., Lugosi, G., Neu, G., Tao, D. (2017b). Algorithmic stability and hypothesis complexity. *In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org*, pp 2159–2167

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). Ssd: Single shot multibox detector. *ECCV* (pp. 21–37). New york: Springer.

Loy, C. C., Lin, D., Ouyang, W., Xiong, Y., Yang, S., Huang, Q., Zhou, D., Xia, W., Li, Q., Luo, P., et al. (2019). Wider face and pedestrian challenge 2018: Methods and results. arXiv preprint arXiv:1902.06854

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

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

Ouyang, W., Wang, X. (2013). Joint deep learning for pedestrian detection. *In: Proceedings of the IEEE international conference on computer vision*, pp 2056–2063

Ouyang, W., Zhou, H., Li, H., Li, Q., Yan, J., & Wang, X. (2017). Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, *40*(8), 1874–1887.

Ouyang, W., Zhou, H., Li, H., Li, Q., Yan, J., & Wang, X. (2018a). Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *T-PAMI*, *40*(8), 1874–1887.

Ouyang, X., Cheng, Y., Jiang, Y., Li, C. L., Zhou, P. (2018b). Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. arXiv preprint arXiv:1804.02047

Park, D., Ramanan, D., & Fowlkes, C. (2010). Multiresolution models for object detection. *ECCV* (pp. 241–254). New york: Springer.

Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., Schiele, B. (2011). Learning people detection models from few training samples. *In: CVPR, IEEE*, pp 1473–1480

Radford, A., Metz, L., Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434

Ran, Y., Weiss, I., Zheng, Q., & Davis, L. S. (2007). Pedestrian detection via periodic motion analysis. *International Journal of Computer Vision*, *71*(2), 143–160.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In: CVPR, IEEE*, pp 779–788

Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *In: NIPS*, pp 91–99

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). New york: Springer.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. *In: CVPR, IEEE*, pp 3234–3243

Sajjadi, M., Javanmardi, M., Tasdizen, T. (2016). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *In: Advances in Neural Information Processing Systems*, pp 1163–1171

Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N. (2018). Deformable gans for pose-based human image generation. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 3408–3416

Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Song, T., Sun, L., Xie, D., Sun, H., Pu, S. (2018). Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. *In: The European Conference on Computer Vision* (ECCV)

Vapnik, V. (2013). *The nature of statistical learning theory*. New york: Springer.

Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H. (2017). Learning to generate long-term future via hierarchical prediction. arXiv preprint arXiv:1704.05831

Viola, P., Jones, M. J., & Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, *63*(2), 153–161.

Vobecky, A., Uricár, M., Hurych, D., Skoviera, R. (2019). Advanced pedestrian dataset augmentation for autonomous driving. *In: ICCV Workshops*, pp 0–0

Wang, X., Shrivastava, A., Gupta, A. (2017). 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

Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C. (2018). Repulsion loss: detecting pedestrians in a crowd. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 7774–7783

Yan, Y., Xu, J., Ni, B., Zhang, W., Yang, X. (2017). Skeleton-aided articulated motion generation. *In: Proceedings of the 2017 ACM on Multimedia Conference, ACM* , pp 199–207

Zanfir, M., Popa, A. I., Zanfir, A., Sminchisescu, C. (2018). Human appearance transfer. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 5391–5399

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. (2016a). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530

Zhang, J., Chen, Z., Tao, D. (2020). Towards high performance human keypoint detection. arXiv preprint arXiv:2002.00537

Zhang, L., Lin, L., Liang, X., & He, K. (2016b). Is faster r-cnn doing well for pedestrian detection? *ECCV* (pp. 443–457). New york: Springer.

Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B. (2016c). How far are we from solving pedestrian detection? *In: CVPR, IEEE*, pp 1259–1267

Zhang, S., Benenson, R., Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. *In: CVPR, IEEE*, vol 1, p 3

Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S. Z. (2018a). Occlusion-aware r-cnn: detecting pedestrians in a crowd. *In: Proceedings of the European Conference on Computer Vision (ECCV)*, pp 637–653

Zhang, S., Yang, J., Schiele, B. (2018b). Occluded pedestrian detection through guided attention in cnns. *In: CVPR, IEEE*, pp 6995–7003

Zheng, Z., Zheng, L., Yang, Y. (2017). 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

Zhu, J.Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *In: ICCV, IEEE*