Pedestrian Detection by Exemplar-Guided Contrastive Learning

Zebin Lin*, Wenjie Pei*, Fanglin Chen*, David Zhang†, Life Fellow, IEEE, and Guangming Lu‡, Member, IEEE

Abstract—Typical methods for pedestrian detection focus on either tackling mutual occlusions between crowded pedestrians, or dealing with the various scales of pedestrians. Detecting pedestrians with substantial appearance diversities such as different pedestrian silhouettes, different viewpoints or different dressing, remains a crucial challenge. Instead of learning each of these diverse pedestrian appearance features individually as most existing methods do, we propose to perform contrastive learning to guide the feature learning in such a way that the semantic distance between pedestrians with different appearances in the learned feature space is minimized to eliminate the appearance diversities, whilst the distance between pedestrians and background is maximized. To facilitate the efficiency and effectiveness of contrastive learning, we construct an exemplar dictionary with representative pedestrian appearances as prior knowledge to construct effective contrastive training pairs and thus guide contrastive learning. Besides, the constructed exemplar dictionary is further leveraged to evaluate the quality of pedestrian proposals during inference by measuring the semantic distance between the proposal and the exemplar dictionary. Extensive experiments on both daytime and nighttime pedestrian detection validate the effectiveness of the proposed method.

Index Terms—Pedestrian detection, contrastive learning.

I. INTRODUCTION

PEDESTRIAN detection is a challenging task in Computer Vision with many important applications, such as video surveillance [1], driving assistance [2] and intelligent robotics [3]. Deep pedestrian detectors [4]–[15], which benefit from excellent feature learning for images by deep neural networks, have achieved great progress in recent years.

Most existing deep pedestrian detectors view pedestrian detection as a particular case of object detection, and are thus designed following the routine object detection methods. A prominent example is Adapted Faster R-CNN [9], which directly adapts Faster R-CNN [16] to pedestrian detection. Based on such pedestrian detection framework of Adapted Faster R-CNN, many methods are proposed to deal with various challenges for pedestrian detection that are not shared in general object detection. To deal with mutual occlusions between adjacent pedestrians in crowded pedestrian detection scenarios, Repulsion loss [17] is designed to maximize the separation between two adjacent pedestrians. To tackle the same challenge of crowded pedestrian detection, OR-CNN [18] performs body division for pedestrian proposals to highlight the visible body parts while suppressing the occluded parts. Another interesting method for addressing the same problem is Case [13], which proposes a count-and-similarity-aware branch to predict the pedestrian number of a proposal.

Another typical challenge of pedestrian detection is how to deal with various scale (size) of pedestrians, which motivates many research works to address it. Typical methods include ALFNet [19] which is designed based on SSD [20] and adopts similar multi-stage predictors as Cascade R-CNN [21], and SAF R-CNN [6] which designs multi-scale detectors to deal with different scales of pedestrians correspondingly.

A crucial challenge of pedestrian detection, which has not been addressed well, is to detect pedestrians with a large amount of appearance diversities, especially in large-scale pedestrian detection scenarios. These appearance diversities are potentially resulted from different body silhouettes, different viewpoints, different dressing, different illumination, etc. A robust pedestrian detector should be insensitive to these appearance diversities but focus on the distinction between pedestrians and background. However, most existing methods perform pedestrian detection following the routine way of object detection, and do not explicitly learn to adapt to (ignore) these intra-class appearance differences. As a result, these methods have to allocate much model capacity for learning to recognize each of these diverse appearances (appeared in training data) as individual positive features for pedestrians, which is hardly generalized to large-scale pedestrian detection.

To address above potential limitation, we propose to perform contrastive learning to guide the feature learning in such a way that two objectives are satisfied: 1) the appearance variations between different pedestrians are ignored in the learned feature space, and 2) the semantic distance in the feature space between pedestrians and background is maximized. To this end, we learn a contrastive feature transformation

Manuscript received 23 September 2021; revised 18 May 2022; accepted 23 June 2022. Date of publication 15 July 2022; date of current version 4 April 2023. This work was supported in part by the NSFC under Grant 62176077, in part by the Shenzhen Key Technical Project under Grant 202202001, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20210324125210025, in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B120120005, in part by the National Natural Science Foundation of China under Grant 62060606 and Grant U2013210, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010306, and in part by the Shenzhen Fundamental Research Program under Grant JCYJ20220818102415032. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaolin Wu. (Corresponding authors: Wenjie Pei; Guangming Lu.)

Zebin Lin, Wenjie Pei, Fanglin Chen, and Guangming Lu are with the Department of Computer Science, Harbin Institute of Technology at Shenzhen, Shenzhen 518057, China (e-mail: 198151113@stu.hit.edu.cn; wenjiecoder@outlook.com; chenfanglin@hit.edu.cn; luguangm@hit.edu.cn).

David Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong at Shenzhen, Shenzhen 518472, China (e-mail: davidzhang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3189803
module by contrastive learning, and embed it into typical pedestrian detection frameworks to project the initial feature space into a new feature space in which above two objectives are achieved. Consequently, the pedestrian diversities are eliminated in the feature space and our method can focus on distinguishing between pedestrians and background, namely a binary-classification task, which is the essential of pedestrian detection.

Typically a robust contrastive learning system demands a large amount of training data to be able to generalize to various positive and negative cases. To facilitate the learning efficiency and effectiveness, we construct an exemplar dictionary for pedestrians to be a representative set covering various appearance variations of pedestrians. The obtained exemplar dictionary is not only used to construct effective training set for contrastive learning, but also leveraged to evaluate the quality of the pedestrian proposals during inference by measuring the semantic similarity between proposals and the exemplar dictionary. Such exemplar-contrastive inference is performed jointly with the typical proposal confidence measurement indicated by classification score to achieve more reliable prediction.

Figure 1 presents several real-world examples for pedestrian detection, in which the learned feature maps are visualized for both the baseline model (Adapted Faster R-CNN) and our EGCL model. Benefiting from the proposed Exemplar-Guided Contrastive Learning framework (EGCL), our model is able to distinguish pedestrians from the background more clearly than the baseline model. To conclude, we make following contributions.

- We construct an exemplar dictionary for pedestrians, which comprises representative pedestrian appearances, to facilitate the efficiency and effectiveness of contrastive learning. Furthermore, the constructed exemplar dictionary is also used to refine the confidence score of proposals during inference, based on the our designed metric for semantic distance between proposals and the exemplar dictionary.
- Based on the exemplar dictionary, we propose an Exemplar-Guided Contrastive Learning framework (EGCL) to guide feature learning such that the intra-class semantic distance between pedestrians in the learned feature space is minimized to ignore appearance diversities while the semantic distance between pedestrians and background is maximized.
- Extensive experiments are conducted on three typical datasets involving both daytime and nighttime pedestrian detection to validate the effectiveness of our method, both in quantitative and qualitative manners.

II. RELATED WORK

In this section, we firstly make a brief review of pedestrian detection. Then we review the related work on pedestrian detection that focuses on tackling occlusions and dealing with different scales of pedestrians, respectively. Finally, we summarize the methods for contrastive learning briefly.

A. Pedestrian Detection

1) Typical Pedestrian Detection: Following the routine object detection methods, pedestrian detectors generally consist of three parts, i.e., proposals generation, feature extraction, and classification and regression. Traditional pedestrian detectors utilize the handcrafted features together with downstream classifiers to detect pedestrians. Thus a lot of handcrafted features are proposed to facilitate the separation between pedestrians and background. Dalal et al. [22] proposes the Histogram of Oriented Gradients (HOG) feature to reflect the pedestrian’s shape and edge information by calculating and integrating the direction and magnitude of the gradient of each pixel. The extracted HOG features are then sent into the downstream classifiers such SVM [22] or AdaBoost [23] to detect pedestrians. Owing to the excellent representation of objects, the HOG feature makes a great improvement in pedestrian detection and its variants such as Aggregated Channel Features (ACF) [24], Integral Channel Features (ICF) [25] and Checkerboards [26] are proposed successively to detect pedestrians.

As the fast development of deep learning, deep pedestrian detection methods [6], [10], [11], [13], [14], [17]–[19] have achieved great progress due to the excellent feature learning capability by deep convolutional networks. Initially,
convolutional neural networks (CNN) are directly used to extracted features to replace the traditional handcrafted features [27], [28]. With the great success of Faster R-CNN [16] on object detection, researchers attempt to adapt the Faster R-CNN into pedestrian detection. To address the issue that the downstream classifier of Faster R-CNN degrades the detection performance, RPN+BF [29] utilizes a boosted forest to substitute for the original classifier of Faster R-CNN and proposes the effective bootstrapping for mining hard negatives. Meanwhile, Adapted Faster R-CNN [9] proposes five improvements to the original Faster R-CNN and exhibits better performance than traditional pedestrian detectors. To eliminate the imbalance between positive samples and negative samples, SDS-RCNN [7] jointly learns pedestrian detection and semantic segmentation by infusing a bounding-box aware semantic segmentation layer into the end of the feature extractor, which enforces the feature focus on the pedestrian's region whilst ignoring the background. Similar to SDS-RCNN, HyperLearner [8] integrates the channel feature generated by a semantic segmentation network into the original feature produced by the standard pedestrian detector. Due to the speed advantage of the one-stage object detector, many pedestrian detectors based on one-stage object detection algorithms such as ALFNet [19] and GDFL [30] are proposed to balance between the detection accuracy and speed of pedestrian detection.

2) Occluded Pedestrian Detection: Occluded pedestrian detection is a challenging task in pedestrian detection due to the information lack of invisible parts of occluded pedestrians and the diversity of occlusions patterns. It aims to solve the problem of mutual occlusions between occluded pedestrians or occlusions caused by background. The first type of pedestrian detectors for solving occlusions mainly utilize the visible parts of occluded pedestrians or divide occluded pedestrian into multiple parts including visible and invisible parts. Bi-box [31] pedestrian detector combines the visible part detection with the full-body detection to simultaneously predict the visible part and full-body part of occluded pedestrian. OR-CNN [18] proposes an aggregation loss to encourage positive proposals to be close to their corresponding ground-truth and focuses on highlighting the visible body parts by suppressing the occluded parts. PCN [10] proposes to predict the score maps of different body parts with a recurrent neural network and designs the context branch for adaptive context selection. Other part-based methods [32], [33] learn a series of part detectors to handle the specific visual patterns of occlusions.

The second type of pedestrian detectors for dealing with occlusions are attention-based methods, which learn robust features with the guidance of the constructed attention maps. Considering that many channels in feature map can be related to different body parts, Zhang et al. [34] proposes a channel-wise branch for reweighting the feature map to highlight the visible parts and suppress occluded parts. MGAN [11] introduces a mask-guided attention branch to greatly eliminate the impact of occluded parts whilst highlighting the visible parts. Other pedestrian detectors for solving occlusions include methods based on feature transformation [35], methods using temporal cue [15].

Both FRCN+A+DT [35] and InterNet [36] are proposed to perform feature transformation to improve the performance of pedestrian detection, which is similar to our method. However, our method substantially differs from these two methods. Our methods differs from FRCN+A+DT [35] in three aspects. Firstly, our method constructs an exemplar dictionary for pedestrians whilst [35] learns one single reference representation for both the pedestrians and the background. Considering the diversity of pedestrian appearance, the exemplar dictionary is able to model such diversities in a more fine-grained manner than the way of learning one single representation. Secondly and most importantly, the whole optimization framework for learning the feature transformation is entirely different between two methods. Our method employs contrastive learning to maximize the margin between positive pairs and negative pairs. In contrast, [35] focuses on minimizing the semantic distance between each sample to the corresponding reference representation. Thirdly, the constructed exemplar dictionary is further leveraged to evaluate the quality of pedestrian proposals during inference by measuring the semantic distance between the proposal and the exemplar dictionary. InterNet [36] is designed for object detection. It learns a representative prototype for each class and minimizes the intra-class distance by the proposed interwiner loss. InterNet differs from our method w.r.t. both the target task and the optimization framework.

3) Multi-Scale Pedestrian Detection: Detecting different scales of objects is another challenge in object detection. As a particular type of object detection, pedestrian detection also needs to deal with different scales of pedestrians, especially the small-scale pedestrians which tend to be blurred and noisy. Many methods follow divide-and-conquer algorithm to detect different scales of pedestrians. For instance, SAF R-CNN [6] trains a large-scale sub-network and a small-scale sub-network to deal with various sizes of pedestrian instances in the image. MS-CNN [37] explores the different depth of feature maps from feature extractor to generate different sizes of proposals, which is followed by a detection network guided by context reasoning. Based on the assumption that different scales of pedestrians bodies can be modeled as 2D Gaussian kernel with various scale variance, TLL [38] designs a unified fully convolutional network to locate the somatic topological line of pedestrians with line annotation for detecting multi-scale pedestrians. Recently, ascribing the poor performance of detect small-scale pedestrians to the problem of inaccurate location, Cao et al. [39] proposes a location bootstrap module for re-weighting the regression loss. The loss of the predicted bounding box far from the corresponding ground-truth is stressed with high weight while the loss of the predicted bounding box near the corresponding ground-truth is ignored with low weight.

Unlike the aforementioned methods for pedestrian detection, we aim to address the challenge of detecting pedestrians with substantial appearance diversities by performing contrastive learning to guide the feature learning to eliminate the appearance diversities in the learned feature space while maximizing the distance between pedestrians and background.
Fig. 2. Architecture of the proposed Exemplar-guided contrastive learning network (EGCL) for pedestrian detection. Built upon the adapted Faster R-CNN, our EGCL performs contrastive learning to learn a feature transformation module $F_t$ in such a way that the semantic distance between pedestrians in the transformed feature space is minimized whilst the distance between pedestrians and background is maximized. The exemplar dictionary is constructed not only for composing high-quality training pairs for contrastive learning, but also for refining the confidence score of predicted proposals by ECI module.

B. Contrastive Learning

Contrastive learning aims to guide the feature learning by minimizing the distance between positive pairs and maximizing the distance between negative pairs in the feature space, typically implemented in the form of contrastive loss [40], [41]. One prominent application of contrastive learning is for the self-supervised learning, which seeks to learn robust feature representation from large-scale unlabeled image dataset using a pretext task [41], [42]. Self-supervised contrastive learning initially aims to eliminate the performance gap between unsupervised learning and supervised learning in image classification by training different pretext tasks. SimCLR [42] learns effective representation by minimizing the discrepancy between differently augmented views of the same data input. MoCo (v1/v2) [41], [43] utilizes a momentum-based moving average of the query encoder to solve the inconsistency of the dictionary keys of negative samples. Recently, considering that previous works for contrastive learning yield sub-optimal performance when transferred to object detection due to the difference between image classification and object detection, DetCo [44] analyses the essential reasons of inconsistency of classification and detection and introduces instance discrimination as a special pretext task for object detection. Meanwhile, apart from computing contrastive loss in high-level features, DetCo also performs contrastive learning on the low-level features for object detection.

Contrastive learning framework typically demands a large amount of positive pairs and negative pairs for training [41], [42], [45]–[47], which requires a lot of computation resources. To alleviate this problem, we propose a novel exemplar-based offline-online training strategy to train our contrastive learning framework for pedestrians detection, which constructs an exemplar dictionary covering representative pedestrian appearances and thereby utilizes the exemplar dictionary to construct high-quality training pairs for efficient contrastive learning.

III. EXEMPLAR-GUIDED CONTRASTIVE LEARNING

We aim to optimize feature learning for pedestrian detection in such a way that the distance between pedestrians with various appearances is minimized whilst maximizing the distance between pedestrians and background. To this end, we propose to perform contrastive learning to optimize the feature learning by viewing pedestrian detection as a binary (pedestrian or background) classification problem. To facilitate the efficiency and effectiveness of contrastive learning, we extract an exemplar dictionary covering representative pedestrian appearances as prior knowledge to guide the contrastive learning. Besides, the pedestrian exemplar dictionary is also leveraged to refine the confidence scores of proposals during inference.

A. Pedestrian Detection Framework

Our proposed contrastive learning framework serves as an auxiliary functional module for optimizing feature learning for pedestrian detection, which can be seamlessly integrated into any existing proposal-based detection framework. As illustrated in Figure 2, we build our contrastive learning module upon the Adapted Faster R-CNN [9] as an instantiation. Adapted Faster R-CNN is a pedestrian detection method designed based on Faster R-CNN [4]. Thus it has similar two-stage modeling process as object detection performed by Faster R-CNN. In the first stage, a Region Proposal Network (RPN) is trained to select a set of high-quality region proposals for potential pedestrians. In the second stage, pedestrian detection is conducted by a classification head and a regression head to predict the class (true or false) of the selected proposals and the corresponding bounding boxes.
(offsets to the ground-truth), respectively. Two stages are both performed in the deep feature space projected by the feature learning head $F_h$, which is typically a pre-trained deep neural networks such as VGG-16 [48] in Adapted Faster R-CNN.

To detect pedestrians with diverse appearance, we learn a contrastive feature transformation module $F_t$ by contrastive learning to project the feature space of the feature learning head $F_h$ into a new feature space. The distance between pedestrians is minimized to eliminate the appearance diversities whilst the distance between pedestrians and background is maximized in this projected feature space. As shown in Figure 2, the contrastive feature transformation module $F_t$ is embedded between the RPN module and the Fast R-CNN module to perform feature transformation for each of selected region proposals by the RPN. Another reasonable position for $F_t$ is between the feature learning head $F_h$ and the RPN, in which case $F_t$ performs feature transformation on the feature maps for the whole image. We conduct experiments to investigate the effect of the position of $F_t$ (before or after RPN) on the performance of pedestrian detection in Section IV-B.

B. Construction of Exemplar Dictionary

The exemplar dictionary is expected to cover full range of appearance diversities of pedestrians. To this end, we perform clustering on the pedestrian images cropped from the training dataset and select the samples closest to each cluster center to compose the exemplar dictionary. Such construction is based on the hypothesis that the pedestrians cropped from a sufficiently large dataset can approximate the pedestrian distribution in real world.

We employ VGG-16 pre-trained on ImageNet [49], which is the same as feature learning head $F_h$ in Figure 2, to extract features for cropped pedestrian images as input features for clustering. Then we perform clustering on the extracted features of all cropped pedestrian images using k-means algorithm [50] and collect the samples closest to each cluster center to obtain the exemplar dictionary $E$:

$$E = \text{k-means}(\{F_h(I_i)\}; K), \quad i = 1, \ldots, N,\tag{1}$$

where $I_i$ is the $i$-th pedestrian image (total $N$ images) cropped from the training set. Here the number of clusters $K$ is a hyper-parameter to control the size of the constructed exemplar dictionary and balance between training efficiency and effectiveness of contrastive learning. Generally, coarse clustering (small $K$) results in a small dictionary with highly representative exemplars, which favors efficient training for contrastive learning but cannot cover full range of pedestrian diversities. In contrast, fine-grained clustering (large $K$) yields a more comprehensive exemplar dictionary but more time is required for contrastive learning to converge. Besides, oversized $K$ may involve low-quality exemplars with low representativeness.

Figure 3 presents a t-SNE [51] map of cropped pedestrian images (blue dots) from the training set of CityPersons [9] and the selected samples (red dots) for constructing the exemplar dictionary. Right: visualized exemplars randomly selected from the exemplar dictionary, exhibit diverse pedestrian appearances.

C. Multi-Level Contrastive Learning

We conduct multi-level contrastive learning to learn the feature transformation module $F_t$, utilizing the obtained exemplar dictionary to construct effective training data for contrastive learning.

1) Contrastive Learning Framework: Figure 4 presents the contrastive learning framework, which is a siamese structure consisting of two parts: the feature transformation module $F_t$ and the projection head $P$. Given a triplet input consisting of three types of image features, namely an exemplar, a positive proposal and $B$ negative proposals, the feature transformation module $F_t$ transforms all three types of input features into a new feature space while the projection head $P$ projects the features into a low-dimensional vector by two fully-connected layers. Then the positive proposal and the exemplar compose the positive training pair while the negative proposals and the exemplar compose $B$ negative training pairs for contrastive learning. Note that both positive proposals and negative proposals are obtained by the RPN module based on a predefined IoU overlap threshold. In our implementation, all negative proposals obtained within a batch are used for composing each of the training triplets, thus the value of $B$ is equal to the number of the negative proposals generated in current batch, which could vary in different batches. Here all input features are extracted from the feature learning head $F_h$ of the backbone network in Figure 2, namely the VGG-16 network. Since the input features are already learned in deep feature space, we design the feature transformation module $F_t$ as a shallow convolutional network, which is composed of three convolutional layers together with activation layers (ReLU in our implementation).

We adopt InfoNCE [52], a widely used contrastive loss function, to supervise our contrastive learning:

$$L_{CL} = \log \frac{\exp\left(\frac{\text{sim}(e, s_{pos})}{\tau}\right)}{\exp\left(\frac{\text{sim}(e, s_{pos})}{\tau}\right) + \sum_{i=1}^{B} \exp\left(\frac{\text{sim}(e, s_{neg}^i)}{\tau}\right)}.$$

\tag{2}
where \( \mathbf{e}, \mathbf{s}_{\text{pos}}, \mathbf{s}_{\text{neg}} \) are the normalized vectorial representations (output from the project head \( \mathcal{P} \)) of the exemplar, the positive proposal and the \( i \)-th negative proposal in a triplet input, respectively. \( \text{sim} \) refers to a kernel function measuring the similarity between paired vectors and we opt for dot-product for \( \text{sim} \) due to its computational efficiency. \( \tau \) is a temperature hyper-parameter [47] to tune the sharpness of the exponential function. Such loss function tries to maximize the similarity of the positive pair while suppressing the similarity values of \( B \) negative pairs. Note that the exemplars \( \mathbf{e} \) in different triplet input can be different since they are randomly selected from the exemplar dictionary.

2) Multi-Level Contrastive Learning Mechanism: A critical issue in object detection is how to deal with different size of objects in a unified framework, which is also challenging in pedestrian detection due to the similar problem formulation. To perform contrastive learning that is robust to different size of pedestrians, we perform multi-level contrastive learning by constructing pyramidal feature pairs using different depth of feature maps for a same training pair, either a positive pair or negative pair. The online contrastive learning is performed after the offline training stage and is conducted jointly with the training of the whole pedestrian detection framework. Thus the parameters of the whole model including the feature learning head \( \mathcal{F}_h \), feature transformation module \( \mathcal{F}_t \) and other modules are jointly optimized under the supervision of the contrastive learning loss and the losses for pedestrian detection (described subsequently). The positive proposals and negative proposals are all from RPN module on the same image that is being processed for pedestrian detection, which are more challenging to distinguish. Hence, the online contrastive learning further improves the performance of feature transformation module by hard-pair training. As shown in Figure 2, the feature transformation module \( \mathcal{F}_t \) is embedded after the RoI Align layer to perform transformation on the features of selected region proposals by the RPN module. Then the transformed proposal features are fed into the Fast R-CNN module to predict the confidence score and the bounding box offsets:

\[
\mathbf{F}_{\text{trans}} = \mathcal{F}_t[\text{RPN}(\mathcal{F}_h(I))],
\]

where \( \mathbf{F}_{\text{trans}} \) is the transformed proposal features by \( \mathcal{F}_t \) and \( I \) is the input image.

D. Collaborative Pedestrian Detection With Exemplar-Contrastive Inference

The pedestrian detection is typically performed by a classification head and a regression head to predict the confidence score and bounding box given the features of a pedestrian proposal, respectively. To better evaluate the quality of the proposal, we further leverage the constructed exemplar dictionary to perform exemplar-contrastive inference to measure the semantic distance between the proposal and the exemplar dictionary.

We measure two kinds of semantic distance between a proposal and the exemplar dictionary: 1) the distance between the proposal and the closest exemplar in the dictionary, 2) average distance from the proposal to the whole exemplar dictionary, which is considered to avoid the overfitting of the proposal to a marginal exemplar (in the cases that the closest exemplar happens to be a marginal exemplar). We employ HNSW [53] algorithm to build a hierarchical graph in a coarse-to-fine manner for the constructed exemplar dictionary (shown in Figure 5), which is used for efficient navigation to locate learning together with the whole pedestrian detection framework, especially focusing on hard-pair contrastive learning for proposals being processed.
in which the parameters of the whole model are trained in an end-to-end manner. Specifically, the whole model is supervised by both the detection loss $L_{\text{det}}$ and contrastive loss $L_{\text{CL}}$: 

$$L = L_{\text{det}} + \alpha L_{\text{CL}},$$

where $\alpha$ is a balancing weight between two terms. The detection loss $L_{\text{det}}$ comprises the losses for RPN stage and the losses for Fast R-CNN stage:

$$L_{\text{det}} = L_{\text{rpn}\_\text{cls}} + L_{\text{rpn}\_\text{reg}} + L_{\text{rcnn}\_\text{cls}} + L_{\text{rcnn}\_\text{reg}}.$$  

Herein, $L_{\text{rpn}\_\text{cls}}$ and $L_{\text{rpn}\_\text{reg}}$ refer to the classification loss and regression loss in RPN stage respectively. Similar notations apply to the $L_{\text{rcnn}\_\text{cls}}$ and $L_{\text{rcnn}\_\text{reg}}$ for Fast R-CNN module.

IV. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of our proposed EGCL model on three benchmark datasets including both daytime and nighttime pedestrian detection scenarios. We first perform ablation study to investigate the effectiveness of each key component of our EGCL. Then we make qualitative comparison between our EGCL and the baseline model (Adapted Faster R-CNN). In the last set of experiments, we compare our model with state-of-the-art methods for pedestrian detection.

A. Experimental Setup

1) Datasets: We evaluate our proposed method on three standard benchmark datasets including CityPersons [9], NightOwls [54] and TJU-DHD-pedestrian [55] involving both daytime and nighttime pedestrian detection scenarios. CityPersons is collected for daytime pedestrian detection. It comprises 2,975, 500 and 1,525 images for training, validation and test, respectively. NightOwls is a nighttime pedestrian dataset, in which all the images are captured in the night and dawn time. It contains 128k images for training, 51k images for validation, and 103k images for test. Covering more complex and diverse scenes than CityPersons and NightOwls, TJU-DHD-pedestrian is a challenging dataset for both daytime and nighttime pedestrian detection. It contains 75,246 images with 373,241 labeled pedestrians, which are mixed of daytime and nighttime images. It has two subsets with different scenes, namely TJU-Ped-campus and TJU-Ped-traffic. Following Pang et al. [55], we conduct experiments on these two subsets separately.

2) Evaluation Metrics: Following the standard pedestrian detection evaluation protocol [56], we choose the log-average Miss Rate over False Positive Per Image (FPPI) with the range of $[10^{-2}, 10^0]$ (denoted as $MR^{-2}$) as the evaluation metric. Lower value of $MR^{-2}$ indicates better performance of pedestrian detection. For experiments on CityPersons dataset, we report evaluation results on 6 subsets according to the visible ratio (in the area of pedestrian bounding boxes) of each pedestrian: $R$ (reasonable subset with visible ratio in $[0.65, 1]$), $HO$ (heavy occlusion subset with visible ratio in $[0.2, 0.65]$), $R+HO$ (reasonable and heavy occlusion subset with visible ratio in $[0.2, 1]$), $Bare$ with visible ratio $[0.9, 1.0]$, $Partial$ with visible ratio $[0.65, 0.9]$ and $Heavy$ with visible ratio $[0, 0.65]$. Following the routine setting [13], [15], the
pedestrians whose bounding box height are at least 50 pixels are taken for evaluation for CityPersons. On NightOwls dataset, following [15], [54] and the official NightOwls evaluation application programming interface (API), we report the evaluation results on the Reasonable subset (non-occluded pedestrians with height $\geq$ 50 pixels), Reasonable_small (non-occluded pedestrians with height between 50 pixels and 75 pixels), Reasonable_occ subsets (occluded pedestrians with height $\geq$ 50 pixels), and All (pedestrians with height $\geq$ 20 pixels). On TJU-DHD-pedestrian dataset, we report evaluation results on R, HO, R+HO and All subsets, which is the same with Pang et al. [55].

3) Implementation Details: For both datasets, we train our framework on the reasonable subset of training data and evaluate on validation sets (test sets of these datasets are not publicly accessible). The backbone network (VGG-16) of our framework is initialized with ImageNet pre-trained model. We use Adam solver [57] as optimizer and the mini-batch size is 60k. Our framework is initialized with ImageNet pre-trained model. The large performance gap between ‘Baseline’ and ‘Baseline+FT’ indicates that embedding the FT module after the RPN module performs better than the other way that before the RPN. We surmise that the FT module placed before the RPN may degenerate the perceptive precision of pedestrians by RPN since the input feature maps of RPN have larger receptive field and thus coarser localization resulted from the convolutional operations in the FT module.

B. Ablation Study

We first perform ablation study to investigate the effectiveness of each key component of our EGCL model including the Feature Transformation module (FT), Offline-Online Contrastive Learning (OOCL) and Exemplar-Contrastive Inference (ECI). Specifically, we conduct experiments which begin with the baseline model (Adapted Faster R-CNN) and then incrementally augment the model with each component of EGCL. Besides, we also conduct experiments to investigate the effectiveness of the constructed exemplar dictionary, Table I and Table II present the experimental results of ablation study on CityPersons and NightOwls, respectively.

1) Comparison With the Baseline (Adapted Faster R-CNN): The experimental results in Table I and Table II show that our EGCL model achieves substantial improvement comparing to the Baseline model (Adapted Faster R-CNN) on all subsets of both CityPersons dataset and NightOwls dataset. Considering the performance on R (Reasonable) subset, which is commonly used for performance comparison, EGCL boosts the performance of pedestrian detection on from 15.40 to 11.79 and from 18.80 to 15.93 (in terms of MR$^{-2}$) on CityPersons and NightOwls respectively.

2) Effect of the Feature Transformation Module (FT): The large performance gap between ‘Baseline’ and ‘Baseline+FT’ on both CityPersons and NightOwls datasets reveals the effectiveness of the feature transformation module of our EGCL. Note that we only perform online contrastive learning to train FT module in the setting of ‘Baseline+FT’. The FT module transforms the initial feature space into a new feature space, in which the semantic distance between pedestrians is minimized while the semantic distance between pedestrians and background is maximized. The comparison between ‘Baseline+FT’ and ‘Baseline+FT+OOCL’ indicates that embedding the FT module after the RPN module performs better than the other way that before the RPN. We surmise that the FT module placed before the RPN may degenerate the perceptive precision of pedestrians by RPN since the input feature maps of RPN have larger receptive field and thus coarser localization resulted from the convolutional operations in the FT module.

3) Effect of Offline-Online Contrastive Learning (OOCL): Comparing the performance between ‘Baseline+FT’ and ‘Baseline+FT+OOCL’, we observe that performing offline contrastive learning before online learning can further boost the performance. It validates the theoretical analysis that offline learning is performed to train the feature transformation module (FT) to be a rough classifier between pedestrians and background while online contrastive learning further improves the performance of FT module.

4) Effect of Exemplar-Contrastive Inference (ECI): The constructed exemplar dictionary is used to not only compose effective training pairs for contrastive learning, but also refine the confidence scores of proposals by proposed exemplar-contrastive inference (ECI). The results in Table I and Table II validate the effectiveness of ECI on both two datasets.

5) Effect of the Constructed Exemplar Dictionary: To investigate the effectiveness of the constructed exemplar dictionary, we compare the performance of contrastive learning using and not using the exemplar dictionary for constructing the training pairs. Figure 6 illustrates the comparison between two ways w.r.t. the loss convergence and performance respectively. Using the constructed exemplar dictionary, the loss of contrastive learning converges faster and reaches a lower value than that

### Table I

| Methods                | Input Scale | R     | HO     | R+HO   | Heavy | Partial | Bare |
|------------------------|-------------|-------|--------|--------|-------|---------|------|
| Baseline [9]           | ×1          | 15.40 | 64.80  | 41.45  | 55.00 | 18.90   | 9.30 |
| Baseline+FT            | ×1          | 13.79 | 55.82  | 31.55  | 54.44 | 14.52   | 8.28 |
| Baseline+FT            | ×1          | 12.81 | 53.58  | 30.57  | 53.21 | 14.15   | 7.28 |
| Baseline+FT+OOCL       | ×1          | 12.42 | 52.25  | 29.37  | 52.47 | 13.21   | 6.78 |
| Baseline+FT+OOCL+ECI   | ×1          | 11.51 | 50.06  | 27.93  | 51.14 | 11.91   | 6.15 |

### Table II

| Methods                | Reasonable | Reasonable_small | Reasonable_occ | All  |
|------------------------|------------|------------------|----------------|------|
| Baseline [9]           | 15.40      | 26.36            | 58.90          | 31.45|
| Baseline+FT+OOCL       | 16.59      | 28.05            | 52.02          | 29.95|
| Baseline+FT+OOCL+ECI   | 15.93      | 26.25            | 50.36          | 28.73|
without exemplar dictionary. Further, exemplar dictionary also empowers contrastive learning to achieve a better performance in terms of $MR^{-2}$, which demonstrates the effectiveness of exemplar dictionary for contrastive learning.

**Effect of the size of exemplar dictionary $K$.** The size of the exemplar dictionary $K$ is a hyper-parameter to be tuned in our experiments. Typically, small $K$ leads to an exemplar dictionary which has highly representative exemplars but cannot cover full range of pedestrian diversities. In contrast, oversized $K$ may involve low-quality exemplars with low representativeness. Figure 7 presents the performance of our EGCL as a function of $K$ on CityPersons and NightOwls datasets respectively. The experiments are conducted under two different settings: 1) the ECI module is unmounted, under which setting the effect of different $K$ on sole contrastive learning is evaluated; 2) the ECI module is activated and the experiments are designed to investigate the effect of different $K$ on both contrastive learning and the exemplar-contrastive inference. We make following observations from the results. 1) Under both settings, the performance of our model is being improved at the beginning as the increase of $K$ and reaches an optimal point, then the performance begins to degrade when provided larger size of exemplar dictionary. These results are consistent with our theoretical analysis. 2) The ECI module consistently improves the performance of our model when setting $K$ to different values, which indicates the robustness of the proposed ECI module. 3) Though the performance of our model fluctuates with varying values of $K$, the performance is always better than the baseline model (Adapted Faster R-CNN), which again reveals the effectiveness of the ECI module.

Since the ‘reasonable’ samples account for the vast majority of pedestrians in the training data and real-world scenarios as well, the exemplar dictionary contains mostly ‘reasonable’ exemplars and only 20% occluded exemplars if it is constructed uniformly based on the training data. However, dealing with the heavy-occluded pedestrians remains a crucial challenge in pedestrian detection. To improve the performance of our EGCL in detecting the heavy-occluded pedestrians, we increase the ratio of occluded samples in the exemplar dictionary by simply replicating the existing occluded exemplars. Table III shows the performance of our EGCL on CityPersons dataset with increasing ratio of occluded samples in the exemplar dictionary. We observe that increasing the ratio of the occluded exemplars indeed improves the performance of our EGCL to detect the heavy-occluded pedestrians (‘HO’ and ‘R+HO’ subsets). However, too large ratio of the occluded exemplars results in the performance degradation of our model on the ‘Reasonable’ subset (R).

**6) Tuning of Hyper-Parameters:** To illustrate the effect of the hyper-parameters involved in our method on the performance, we report the performance of our method with different settings of hyper-parameters. To be specific, Figure 8 visualizes the distribution map of performance as a function of $w_i, i = 3, ..., 5$ in Equation 3 when performing grid search for 3 parameters. Note that $w_2$ is fixed to be 1 to tune other balancing weights. Figure 9 shows the effect of $\alpha$ in Equation 8 on the performance while Figure 10 presents the...
the effectiveness of our EGCL, we make a qualitative comparison between our EGCL and the baseline model (Adapted Faster R-CNN) by visualizing the feature maps of both models. Figure 11 visualizes both the feature maps of the whole images (C5 block of feature learning) and that for cropped region proposals (RoIs). Our EGCL are also visualized in Figure 11 for reference. An interesting observation is that our model is able to correctly distinguish between positive proposals (IoU > 0.6) and hard negative proposals (IoU < 0.4). Figure 13 presents t-SNE maps for the baseline model and our model (including two cases: before and after the feature transformation module (FT)). The t-SNE map of RoI features after the FT module by our EGCL shows more separability than other two t-SNE maps.

2) Separability of RoI Features: We further apply t-SNE [51] to RoI features by different models to compare their capability to distinguish between positive proposals (IoU > 0.6) and hard negative proposals (IoU < 0.4). Figure 13 presents t-SNE maps for the baseline model and our model. We first present the detection results of both models on three examples that are randomly selected from the training set in Figure 12 (a). The results show that our model is able to detect all pedestrians precisely, even those with quite small size (in the second sample) or occluded heavily (in the third sample) which are missed by the baseline model. Besides, the baseline falsely captures the reflection of the pedestrian in the mirror (in the first sample) while our model can discriminate the reflection from real pedestrians.

C. Qualitative Study on Contrastive Learning

1) Visualization of Learned Feature Maps: To obtain more insights into the effectiveness of our EGCL, we make a qualitative comparison between our EGCL and the baseline model (Adapted Faster R-CNN) by visualizing the feature maps of both models. Figure 11 visualizes both the feature maps for the whole images (C5 block of feature learning head $F_k$) and that for cropped region proposals (RoIs). Our EGCL is able to detect pedestrians more precisely since it minimizes the semantic difference between pedestrians and maximizes the distance between pedestrians and background in the feature space. Besides, the untransformed feature maps for RoIs before the feature transformation module of our EGCL are also visualized in Figure 11 for reference. An interesting observation is that our model is able to correctly deny the confusing (hard) negative sample in the last row of RoI visualization, which tends to be falsely recognized as a pedestrian by the baseline and the untransformed feature maps of our model.

D. Comparison With State-of-the-Art Methods

1) CityPersons Dataset: To evaluate the performance of our EGCL on daytime pedestrian detection, we compare our EGCL model with other state-of-the-art methods for pedestrian detection on CityPersons dataset. These methods include F.RCNN+ATT-vbb [34], F.RCNN+ATT-part [34], Adapted Faster R-CNN [9], OR-CNN [18], Adaptive-NMS [58], MGAN [11], HGPD [59], TLL [38], $R^2$NMS [14], TLL+MRF [38], Replusion loss [17], ALFNet [19], NOH-NMS [61], FRCN+A+DT [35], Case [13] and Crowd-Det [60]. We also adapt InterNet [36] from object detection.
Fig. 11. Visualization of the feature maps learned by the Baseline model (Adapted Faster R-CNN) and by our EGCL model for randomly selected samples from CityPersons [9] validation dataset. Left: the feature maps of whole images (C5 block of the feature learning head $F_h$ before RPN module) are visualized for both the baseline model and our EGCL. Right: the feature maps (resized to $7 \times 7$) for cropped region proposals (RoIs by RPN module) are visualized for both models. Note that the last sample of RoI visualization is a hard negative sample which tends to be falsely recognized as a pedestrian by the baseline.

Fig. 12. Visualization of detection results by our model and the baseline. (a) Our model is able to detect all pedestrians correctly whilst the baseline cannot recognize the pedestrians with small-scale or heavily occluded pedestrians. (b) Two challenging examples on which both our model and the baseline fail to detect pedestrians correctly.

Fig. 13. The t-SNE map of RoI features by different models, in which the blue dots and the red dots refer to the positive proposals ($IoU > 0.6$) and negative proposals ($IoU < 0.4$). Left and Middle: RoI features before and after the feature transformation module (FT) of our trained EGCL respectively; Right: RoI features of the baseline model.

2) NightOwls Dataset: Detecting pedestrians in nighttime is more challenging than in daytime due to difficult discrimination between pedestrians and background in blurred and low-contrast circumstances. To validate the effectiveness of our model in nighttime pedestrian detection, we compare our EGCL model with state-of-the-art pedestrian detectors on NightOwls dataset. These methods include ACF [24], Checkerboards [26], Vanilla Faster R-CNN [16], Adapted...
TABLE V

| Methods               | Backbone | Input Scale | R   | HO  | R+HO | Heavy | Partial | Bare |
|-----------------------|----------|-------------|-----|-----|------|-------|---------|------|
| Faster R-CNN [34]     | VGG-16   | x1          | 16.4| 57.3| -    | -     | -       | -    |
| FRCNN-ATT-vbb [34]    | VGG-16   | x1          | 16.0| 56.7| 38.2 | -     | -       | -    |
| Adapted FasterRCNN [9]| VGG-16   | x1          | 15.4| 64.8| 41.5 | 55.0  | 18.9    | 9.3  |
| OR-CNN [18]           | VGG-16   | x1          | 12.8| 55.7| -    | 55.7  | 15.3    | 6.7  |
| Adaptive-NMS [58]     | VGG-16   | x1          | 11.9| 55.2| -    | 55.2  | 12.6    | 6.2  |
| MGAN [11]             | VGG-16   | x1          | 11.5| 51.7| -    | -     | -       | -    |
| HGPD [59]             | VGG-16   | x1          | 11.3| 51.7| -    | -     | -       | -    |
| R²NMS [14]            | VGG-16   | x1          | 11.1| 53.3| -    | -     | -       | -    |
| Case [13]             | VGG-16   | x1          | 11.0| 50.3| -    | -     | -       | -    |
| EGCL (ours)           | VGG-16   | x1          | 11.5| 50.0| 27.9 | 51.1  | 11.9    | 6.1  |

The results in Table VI show that our model outperforms other methods, which reveals the advantage of our model in nighttime detection over other methods.

3) TJU-DHD-Pedestrian Dataset: In the last set of experiments, we compare our EGCL with other methods on TJU-DHD-pedestrian dataset, which is mixed with daytime and nighttime images covering more challenging scenes for pedestrian detection. We compare our EGCL with the state-of-the-art methods that have been evaluated previously faster than R-CNN [9], RPN+BF [29], SDS-RCNN [7] and recent TFAN [15].

TABLE VI

Performance (in terms of MR^2) of Our EGCL and Other Methods on CityPersons Validation Subset (Lower Is Better)

| Methods               | Backbone | Reasonable |
|-----------------------|----------|------------|
| ACF [24]              | -        | 51.68      |
| Checkerboards [26]    | -        | 39.67      |
| Vanilla Faster R-CNN [16] | VGG-16 | 20.00      |
| Adapted Faster R-CNN [9] | VGG-16 | 18.81      |
| RPN+BF [29]           | VGG-16   | 23.26      |
| SDS-RCNN [7]          | VGG-16   | 17.80      |
| TFAN [15]             | ResNet-50| 16.50      |
| EGCL (ours)           | VGG-16   | 15.93      |

TABLE VII

Performance (in terms of MR^2) of Our EGCL and Other Methods on TJU-DHD-Pedestrian Dataset Including Two Sub-Datasets (Lower Is Better)

| Subset              | Methods   | R   | HO  | R+HO | ALL   |
|---------------------|-----------|-----|-----|------|-------|
| TJU-Ped-campus      | RetinaNet [62] | 34.73| 71.31| 42.26| 44.34 |
|                     | FCOS [63]  | 31.89| 69.04| 39.38| 41.62 |
|                     | CrowdDet [60] | 27.92| 67.52| 35.67| 38.08 |
|                     | CrowdDet+EGCL (ours) | 25.73| 66.38| 33.65| 35.90 |
|                     | RetinaNet [62] | 23.89| 61.60| 28.45| 41.40 |
|                     | FCOS [63]  | 24.35| 63.73| 28.86| 40.02 |
|                     | CrowdDet [60] | 22.30| 60.30| 26.71| 37.78 |
|                     | CrowdDet+EGCL (ours) | 20.82| 61.22| 25.28| 36.94 |
|                     | RetinaNet [62] | 19.73| 60.05| 24.19| 35.76 |

The results in Table VII show that our model achieves the best results on all subsets of both ‘campus’ and ‘traffic’ sub-datasets, which validates the effectiveness of our method. Note that only...
methods with reported detection results on this dataset are listed. Here CrowdDet is used as the baseline for our model to have a fair comparison. Our EGCL performs better than CrowdDet on all metrics.

V. CONCLUSION

In this paper, we have presented the Exemplar-Guided Contrastive Learing (EGCL) model for pedestrian detection. EGCL learns a feature transformation module by contrastive learning to project the initial feature space into a new feature space, in which the semantic distance between pedestrians is minimized to eliminate the appearance diversities of pedestrians while the semantic distance between pedestrians and background is maximized. Extensive experiments on both daytime and nighttime pedestrian detection validate the effectiveness of the proposed EGCL.

REFERENCES

[1] J. C. Nascimento and J. S. Marques, “Performance evaluation of object detection algorithms for video surveillance,” IEEE Trans. Multimedia, vol. 8, no. 4, pp. 761–774, Aug. 2006.
[2] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 3354–3361.
[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” Int. J. Robot. Res., vol. 32, no. 11, pp. 1231–1237, 2013.
[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” 2015, arXiv:1506.01497.
[5] X. Du, M. El-Khamy, J. Lee, and L. Davis, “Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2017, pp. 953–961.
[6] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, “Scale-aware fast R-CNN for pedestrian detection,” IEEE Trans. Multimedia, vol. 20, no. 4, pp. 985–996, Aug. 2017.
[7] G. Brazil, X. Yin, and X. Liu, “Illuminating pedestrians via simultaneous detection and segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4950–4959.
[8] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, “What can help pedestrian detection?” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3127–3136.
[9] S. Zhang, R. Benenson, and B. Schiele, “CityPersons: A diverse dataset for pedestrian detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3213–3221.
[10] S. Wang, J. Cheng, H. Liu, and M. Tang, “PCN: Part and context information for pedestrian detection with CNNs,” 2018, arXiv:1804.04483.
[11] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, “Mask-guided attention network for occluded pedestrian detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4967–4975.
[12] X. Song, K. Zhao, W.-S. C. H. Zhang, and J. Guo, “Progressive refinement network for occluded pedestrian detection,” in Proc. ECCV, vol. 7, 2020, p. 9.
[13] J. Xie et al., “Count- and similarity-aware R-CNN for pedestrian detection,” in Proc. ECCV, Cham, Switzerland: Springer, 2020, pp. 88–104.
[14] X. Huang, Z. Ge, Z. Jie, and O. Yoshe, “NMS by representative region: Towards crowded pedestrian detection by proposal pairing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10750–10759.
[15] J. Wu, C. Zhou, M. Yang, Q. Zhang, Y. Li, and J. Yuan, “Temporal-context enhanced detection of heavily occluded pedestrians,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13430–13439.
[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2016.
[17] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, “Repulsion loss: Detecting pedestrians in a crowd,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7774–7783.
Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” 2019, arXiv:1906.05849.

Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3733–3742.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Math. Statist., Prob., Oakland, CA, USA, Jun. 1967, vol. 1, pp. 281–297.

L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 2579–2605, 2008.

A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, arXiv:1807.03748.

Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 824–836, Apr. 2018.

Lei Neumann et al., “NightOwls: A pedestrians at night dataset,” in Proc. ACCV Cham, Switzerland: Springer, 2018, pp. 691–705.

Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, “TJU-DHD: A diverse high-resolution dataset for object detection,” IEEE Trans. Image Process., vol. 30, pp. 207–219, 2020.

P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 743–761, Apr. 2011.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

S. Liu, D. Huang, and Y. Wang, “Adaptive NMS: Refining pedestrian detection in a crowd,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6459–6468.

G. Li, J. Li, S. Zhang, and J. Yang, “Learning hierarchical graph for occluded pedestrian detection,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1597–1605.

X. Chu, A. Zheng, X. Zhang, and J. Sun, “Detection in crowded scenes: One proposal, multiple predictions,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12214–12223.

P. Zhou et al., “NOH-NMS: Improving pedestrian detection by nearby objects hallucination,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1967–1975.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9627–9636.

T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.

Zebin Lin received the B.S. degree in network engineering from South China Normal University in 2018 and the M.S. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, in 2022. His current research interests include object detection, medical diagnosis, pattern recognition, and deep learning.

Wenjie Pei received the Ph.D. degree from the Delft University of Technology, The Netherlands, in 2018, working with Dr. Laurens van der Maaten and Dr. David Tax. He is currently an Assistant Professor with the Harbin Institute of Technology, Shenzhen, China. Before joining the Harbin Institute of Technology, he was a Senior Researcher in computer vision with the Tencent YouTu X-Laboratory. In 2016, he was a Visiting Scholar with Carnegie Mellon University. His research interests lie in computer vision and pattern recognition, including sequence modeling, deep learning, and video/image captioning.

Fanglin Chen received the B.S. degree in control science and engineering and the B.S. degree in applied mathematics from Xi’an Jiaotong University, Xi’an, China, in 2006, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2011. He is currently a Professor with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. His current research areas include computer vision, pattern recognition, and image processing. He received the Academic Newcomer Award (Ministry of Education, China) in 2010, and the First Group of Robot Vision Challenge (ImageCLEF) in 2014.

David Zhang (Life Fellow, IEEE) received the B.Sc. degree in computer science from Peking University, the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT) in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 1994. From 1986 to 1988, he was a Postdoctoral Fellow with Tsinghua University. Then, he was an Associate Professor with the Academia Sinica, Beijing. He is currently a Professor with The Chinese University of Hong Kong (Shenzhen). He also serves as a Visiting Chair Professor with Tsinghua University and HIT, and an Adjunct Professor with Shanghai Jiao Tong University, Peking University, National University of Defense Technology, and the University of Waterloo. He is the Founder and the Editor-in-Chief of the International Journal of Image and Graphics (IJIG), the Book Editor of International Series on Biometrics (KISB) (Springer), the Organizer of the First International Conference on Biometrics Authentication (ICBA), and an Associate Editor of more than ten international journals, including IEEE TRANSACTIONS. He has published over 20 monographs, 400 international journal articles, and 40 patents from USA/Japan/Hong Kong/China. He is a Croucher Senior Research Fellow, the Distinguished Speaker of the IEEE Computer Society, and a fellow of IAPR.

Guanming Lu (Member, IEEE) received the bachelor’s degree in electrical engineering, the master’s degree in control theory and control engineering, and the Ph.D. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1998, 2000, and 2005, respectively. From 2005 to 2007, he was a Postdoctoral Fellow with Tsinghua University. He is currently a Professor with the Harbin Institute of Technology, Shenzhen, China. He has published over 100 technical papers at some international journals and conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, NeurIPS, CVPR, AAAI, and IJCAI. His research interests include computer vision, pattern recognition, and machine learning.