applied sciences

MDPI

Article

A Shallow–Deep Feature Fusion Method for Pedestrian Detection

Daxue Liu [1,*], Kai Zang [2] and Jifeng Shen [3]

[1] College of Intelligence Science, National University of Defense Technology, Changsha 410073, China
[2] School of Automation, Southeast University, Nanjing 210096, China; kaizang126@126.com
[3] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China; shenjifeng@ujs.edu.cn
* Correspondence: daxueliu@nudt.edu.cn; Tel.: +86-731-8700-5430

Abstract: In this paper, a shallow–deep feature fusion (SDFF) method is developed for pedestrian detection. Firstly, we propose a shallow feature-based method under the ACF framework of pedestrian detection. More precisely, improved Haar-like templates with Local FDA learning are used to filter the channel maps of ACF such that these Haar-like features are able to improve the discriminative power and therefore enhance the detection performance. The proposed shallow feature is also referred to as weighted subset-haar-like feature. It is efficient in pedestrian detection with a high recall rate and precise localization. Secondly, the proposed shallow feature-based detection method operates as a region proposal. A classifier equipped with ResNet is then used to refine the region proposals to judge whether each region contains a pedestrian or not. The extensive experiments evaluated on INRIA, Caltech, and TUD-Brussel datasets show that SDFF is an effective and efficient method for pedestrian detection.

Keywords: feature extraction; ACF; Haar-like feature; Local FDA; ResNet; pedestrian detection

check for updates

Citation: Liu, D.; Zang, K.; Shen, J. A Shallow–Deep Feature Fusion Method for Pedestrian Detection. *Appl. Sci.* **2021**, *11*, 9202. https://doi.org/10.3390/app11199202

Academic Editors: Antonio Fernández-Caballero, Byung-Gyu Kim and Hugo Pedro Proença

Received: 1 September 2021
Accepted: 27 September 2021
Published: 3 October 2021

1. Introduction

Pedestrian detection is a crucial research topic in pattern recognition and computer vision since it can be widely applied in video surveillance, action analysis, and advanced driver assistance systems (ADAS) [1,2]. The performance of pedestrian detection is still vulnerable to massive challenges in real-world applications due to illumination variances, pose changes, occlusion, and human deformation.

Pedestrian detection is usually viewed as a particular object-detection problem, i.e., rigid or half-rigid object detection [3–5]. It is significantly different from human detection with respect to the image resolution, shooting angle, and posture deformation, where complicated models are usually used, such as multi-view modeling and explicit geometric modeling [6], for example. In general, the pedestrians with similar aspect ratio and shape appearance are prone to several challenges in terms of size, occlusion and illumination that are usually caused by the camera shooting angle and the specific constraints of the scene. For instance, pedestrians distant from the camera contain weak appearance information and thus are often incorrectly recognized as the surrounding long-thin objects due to the self-occlusion in body profile and crowd.

In this paper, a shallow–deep feature fusion method (SDFF) for pedestrian detection is proposed. In the latter, a shallow feature method is used to rapidly produce pedestrian proposals with high recall rate followed by a deep feature method aiming to efficiently remove the false positives. Firstly, SDFF generates precise pedestrian proposals with high recall rate using a shallow feature method, such asthe previously proposed method in [7]. SDFF then efficiently removes the false positives using the ResNet deep feature-based method [8].

The contributions of this paper can be summarized as follows:

(1) Under ACF framework, a shallow feature-representation-based pedestrian detector is proposed. Firstly, Haar-like templates are utilized to filter the channel maps in order

to improve the feature discrimination. The local FDA learning is then used to distill the discrimination. The shallow feature representation-based pedestrian detection works not only independently but also on region proposals.

In Figure 1, Ω is a multi-channel feature extraction operation, Σ is a max pooling operation, and $(:)$ is a vectorize operation.

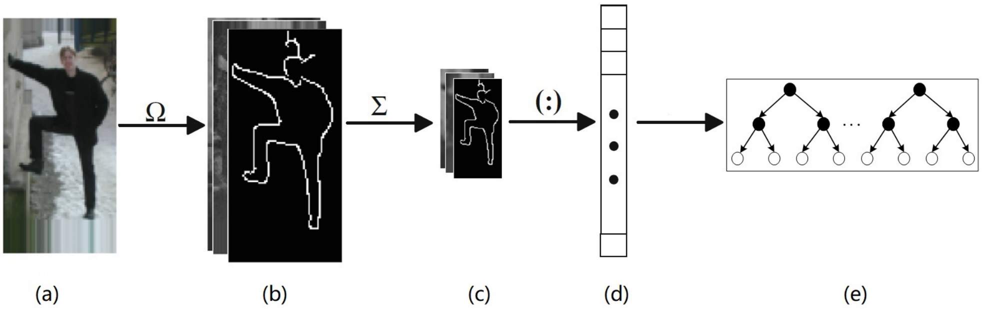

Figure 1. The flowchart of ACF. (**a**) Original color image (**b**) Extended image channels (**c**) Aggregate channels (**d**) Feature vector (**e**) Soft cascade.

(2) Under region proposal and refinement framework, ResNet is used to refine the region proposals generated from the shallow feature-representation-based pedestrian detection. Thus, a shallow feature and deep feature fusion (SDFF) for pedestrian detection is built.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 details the proposed method. Section 4 presents the extensive experimental evaluations used to validate the performance of the proposed method. Finally, conclusions are drawn in Section 5.

2. Related Work

In the past decades, massive efforts were devoted to promoting pedestrian detection. Existing methods can be grouped into two categories: (1) pedestrian detection via sliding window; (2) pedestrian detection via region proposal.

2.1. Pedestrian Detection via Sliding Window

Before deep learning, pedestrian detection could be viewed as a binary classification task that identifies whether the given image window is a pedestrian. The earlier work scans the whole image with a sliding window and classifies the candidate window. More precisely, pedestrian detection based on a sliding window includes two steps: feature extraction and classification.

Firstly, Wavelet [9], HOG [5], Haar wavelet [10], and contour [11] are the classical hand-crafted descriptors for object detection. The first order differential of the image based on HOG is very robust to the illumination changes. However, HOG performs poorly when the occlusion and complex background are present. Haar wavelet is computationally efficient since it can be calculated through integral images. However, it contains some discriminative information because of its simple structure. Afterwards, several improved studies [12,13] are developed under the Adaboost framework [10]. On the other hand, pedestrian detection based on modeling the relationship of the human parts using HOG [6,14] achieves great success under the conditions of pose changes and occlusion. In order to detect generic objects with large deformation, the mixture strategies of different detection methods are often employed [15]. For instance, Piotr et al. [16] combined HOG and LUV color features to build integral channel features (ICF) for object detection. ICF can efficiently represent the pedestrian in the color image and can be utilized to detect pedestrian with 30+ FPS for 640 × 480 images. Following ICF, Piotr et al. further proposed a more efficient aggregated channel features (ACF) framework for pedestrian detection, shown in Figure 1. ACF inherits the ten channel maps from ICF. Firstly, the probe image I is smoothed, and then the ten channel maps are calculated as shown in Figure 1b. Afterwards, each channel is down-sampled to one-quarter size as shown in Figure 1c and concatenated into column feature vectors as shown in Figure 1d. Finally, Adaboost with decision trees is trained as the classifier over these features in order to distinguish the object from the others as shown in Figure 1e.

Furthermore, the progress made in pedestrian detection demonstrates that filtering on the channel maps of ACF can further enhance the discriminative power of the feature and therefore improve the pedestrian detection performance [17–20]. In addition, several auxiliary visual clues exemplified by context information [21], semantic information [22,23], ad optical flow information [24] can be used as supplementary contents to further enhance the detection performance.

2.2. Region Proposal via Pedestrian Detection

On the contrary, region proposal-based pedestrian detection methods can generate a small number of candidate windows. A large number of region proposal methods were developed, e.g., BING [25], EdgeBox [26], Selective Search [27], Objectness [28], and CPMC [29]. More complex feature representation methods [30,31] were also used to enhance the detection performance without affecting the detection efficiency. However, the region proposal methods usually produce inaccurate bounding boxes, and thus the bounding box regression is usually performed subsequently for refinement.

Deep convolutional neural networks (DCNN) [32–34] demonstrated a high efficiency in image classification [33] and object detection [35,36]. They integrate the object classification and the object bounding box regression into a unified framework using an end-to-end learning strategy, e.g., Faster RCNN [35], SSD [37], and YOLO [38]. Object detection using DCNN is usually divided into two classes: two-stage methods and one-stage methods. As a representative method of the former class, Faster RCNN utilizes Region Proposal Network (RPN) as region proposal generation module and adopts Fast RCNN to accurately classify objects and refine the bounding box. On the contrary, SSD and YOLO are one-stage object detection methods. SSD utilizes multiple extra convolutional layers to output the class labels and the bounding box of the object in order to improve the performance. YOLO utilizes the last convolutional layer to generate the class label and the bounding box of the object, resulting in the fast deep feature-based generic object detection method. Cheng et al. [39] analyzed the failure cases of the region-based object detectors and explored their potentials in accuracy improvement. More recently, Zhang et al. [40] used RPN in order to to perform region proposal and re-scoring using a adaboost with decision trees to eliminate false positives. Cai et al. [41] proposed to detect pedestrian with multiple sizes using different levels of deep feature maps such that it is easy to find the small-size pedestrians. Liu et al. [42] proposed RFBNet for object detection, in which the later convolutional layers of SSD are simply replaced with a Receptive Field Block (RFB) in order to to improve the features discriminability, and robustness. These methods usually depend on GPU and therefore they have a high computational cost.

3. The Proposed Work

3.1. The Proposed Pedestrian Detection Framework

This section presents the proposed shallow–deep feature fusion method (SDFF) for pedestrian detection. The method consists of a two-stage cascade structure that involves a region proposal generation module and a pedestrian refining module. In the first stage, shallow feature-based pedestrian proposals are generated. The Haar template is then used to filter the channel maps of ACF in order to to build a Subset-Haar-Like feature such that the column vector can be built to improve the discriminative information. Afterwards, the local Fisher discriminant analysis (LFDA) [43] is used to further improve the discrimination of the column vectors. In the second stage, a classifier is built with ResNet as a backbone in order to classify whether the pedestrian candidates are true pedestrians or not.

In Figure 2, filter is a filtering operation with multiple Haar-like templates. LFDA is a local feature extraction method.

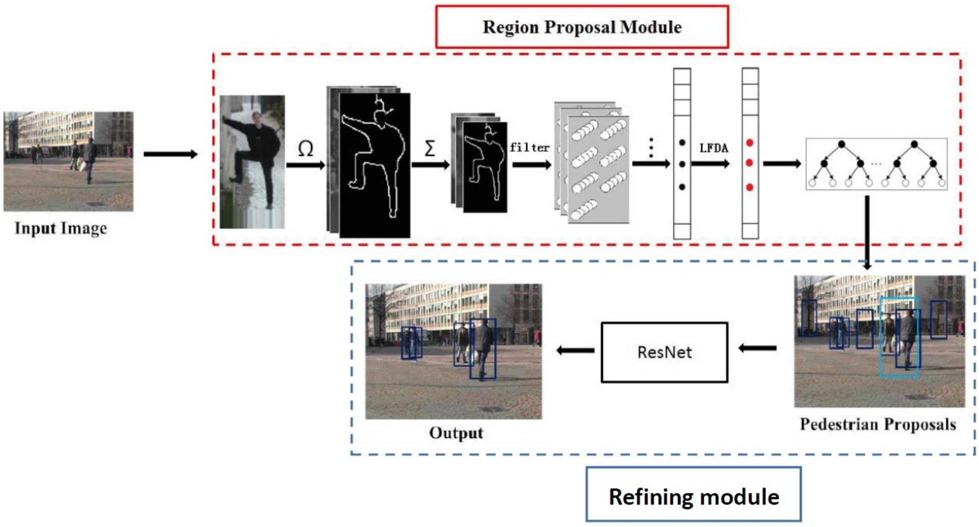

Figure 2. The flowchart of SDFF for pedestrian detection.

3.2. *Shallow Feature-Region Proposal Generation*

For subset-Haar-like features, pedestrians with different standing postures usually share very similar structures, e.g., aspect ratio, contour, and edge. They usually contain two characteristics: the strong edge and the significant structure. Herein, seven types of Haar-like templates are utilized to filter the channel maps of ACF and build the intermediate feature maps, referred to as Subset-Haar-like features, as illustrated in Figure 3, where T1, ..., T7 are seven templates.

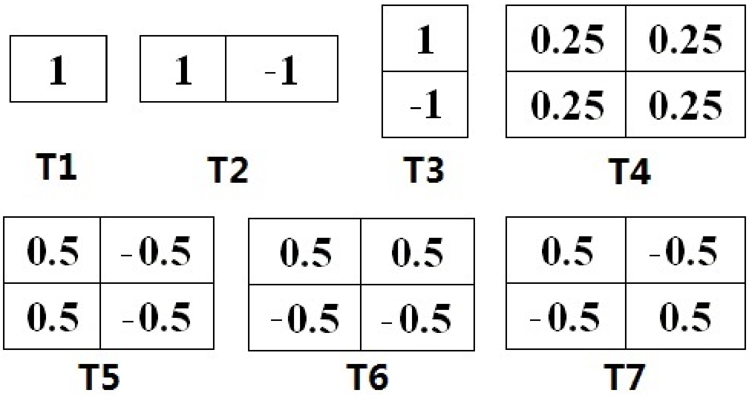

Figure 3. Seven Haar-like templates.

The algorithm of Subset-Haar-like feature generation performs as follows:

Firstly, all the training samples are cropped and resized to 128 × 64.

The ten channel maps of all the training samples that are similar to ACF are then calculated and down-sampled with the shrink parameter, yielding each channel map of 32 × 16.

Afterwards, the ten channels maps are filtered with the seven Haar-Like templates in order to build Subset-Haar-like feature maps.

Finally, all the Subset-Haar-like feature maps are concatenated to generate the final feature vectors.

Seven Haar-like templates are used to filter the ten channels maps with a resolution of 32 × 16. The dimensions of the seven Subset-Haar-like feature maps are shown in Table 1. Thus, the total dimension of the ten Subset-Haar-like feature maps of ACF is 3348 × 10, while the dimensions of ten channel maps of ACF is 32 × 16 × 10. Although it significantly increases the feature dimension of ACF, Subset-Haar-Like feature maps can be efficiently calculated. More specifically, the pedestrian proposal method with ACF and subset-Haar-like features can reach 35 and 32 FPS, which is very efficient to calculate.

Table 1. The dimensions of the seven Subset-Haar-like features.

Template#	Feature Map Size	Feature Dimension
T1	32 × 16	512
T2	32 × 15	480
T3	31 × 16	496
T4	31 × 15	465
T5	31 × 15	465
T6	31 × 15	465
T7	31 × 15	465
Total	/	3348

For LFDA-based feature learning, the subset-Haar-like feature is generated by filtering the Haar-like template on the channel maps of ACF. Since the structure of the Haar-like templates is simple and computationally efficient in terms of implementation, the discriminating power can be further improved by subspace learning, e.g., Fisher discriminant analysis (FDA) [44] and Local FDA (LFDA). Herein, we rebuild the subset-Haar-like feature by learning the combining coefficients using LFDA.

Given the training samples $x_i(i = 1, 2, ..., n)$ with the corresponding class label y_i, the number of the lth class samples is n_l. LFDA aims to seek the projection matrix by maximizing the ratio between the local between-class scatter and the local within-class scatter. The local between-class scatter matrix S_{lb} and the local within-class scatter matrix S_{lw} are expressed as:

$$S_{lb} = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{b}(x_i - x_j)(x_i - x_j)^T \tag{1}$$

$$S_{lw} = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{w}(x_i - x_j)(x_i - x_j)^T \tag{2}$$

$$W_{ij}^{b} = \begin{cases} A_{ij}(\frac{1}{n} - \frac{1}{n_j}), & y_i = y_j \\ \frac{1}{n}, & \text{else} \end{cases} \tag{3}$$

$$W_{ij}^{w} = \begin{cases} A_{ij}\frac{1}{n_l}, & y_i = y_j \\ 0, & \text{else} \end{cases} \tag{4}$$

$$A_{ij} = \exp\left(\frac{||x_i - x_j||^2}{\sigma_i \sigma_j}\right) \tag{5}$$

where W_{ij}^{b} and W_{ij}^{w} are respectively the affine weights in between-class samples and in within-class samples, A_{ij} is the affinity between x_i and x_j and σ_i is the local scaling of the nearest neighbors of x_i.

Finally, the projection matrix P is calculated as:

$$P = \arg\max \frac{tr(P^T S_{lb} P)}{tr(P^T S_{lw} P)} \tag{6}$$

Given the Subset-Haar-like feature, n features are randomly selected from each channel map and fused into a new feature with the weight from LFDA. Thus, 33,480 features can be generated in all the ten channel maps with 3348 new features in each channel map.

The procedure of FLDA-based feature learning on the Subset-Haar-like feature is summarized as follows. Firstly, both positive and negative samples are collected. The ten channel maps of all the samples are then built. Afterwards, the subset-Haar-like features of all the samples are obtained by filtering on the ten channel maps. Finally, the weighted Subset-Haar-like feature of all the samples is built using LFDA.

3.3. Deep Feature Based Pedestrian Refining

The latest research [45,46] shows that the depth of the network has a crucial influence on the detection performance. Generally, a deeper network leads to a degraded performance and saturated accuracy. For instance, He et al. [8] proposed the Deep Residual Network. It is the most important characteristic of ResNet to solve the problem of DCNN degradation by introducing residual learning and fast identity connection. The theoretical analysis and experimental evaluations demonstrated that ResNet has higher classification accuracy than other similar DCNN on PASCAL VOC 2007, 2012, and COCO datasets.

The current mainstream learning framework for pedestrian detection based on DCNN comprises two steps: region proposal generation module and classification module. Inspired by the approach of [17], which combines traditional machine learning and DCNN for pedestrian detection, we further utilize ResNet to refine the region proposals from the weighted Subset-Haar-like feature module and remove the false positives. The algorithm of pedestrian refining based on ResNet performs as follows.

Firstly, the set of positive and negative samples are built. More precisely, the weighted subset-Haar-like feature-based region proposal module is used to detect training the positive samples and crop out the ground truth region and detection-bounding boxes with IoU > 0.7, producing the images used as the training positive samples. Similarly, detected bounding boxes with IoU < 0.3 are used as the training negative samples.

The ResNet152 deep model is then trained. The training samples are used to fine-tune the ResNet152 pre-trained on the PASCAL VOC 2007 database.

Afterwards, the pedestrian proposals are refined. Finally the region proposals derived from the weighted subset-Haar-like feature based region proposal module are delivered into the final fine-tuned ResNet152 model for classification.

4. Experiments

4.1. The Dataset and Setup

The experiments were conducted on the INRIA, Caltech, and TUD-Brussel datasets to evaluate the different methods.

INRIA [47]: The training set includes 2416 pedestrians with annotations assembled from 614 positive images and 1218 non-pedestrian images.The testing set includes 1132 pedestrian with annotations from 288 images and 453 non-pedestrian images.

Caltech [48]: The training set includes 6325 pedestrians with annotations assembled from 4250 images (set00-set05), while the testing set includes 5051 pedestrians with annotations collected from 4024 images (set06-set10).

TUD-Brussel [49]: The training set contains 1092 image pairs with 1776 annotated pedestrians and 192 image pairs in the negative set. It also contains 26 additional image pairs with 183 pedestrians used for hard data mining, and 508 image pairs with 1326 annotated pedestrians are used for testing.

The positive samples were cropped and resized to 128×64. For the INRIA dataset, the number of the positive samples was extended to 2474. Since the pedestrian heights on INRIA database vary from 60 to 780 pixels, we do not need to upsample the test images in the testing phase. On the Caltech dataset, the number of the positive samples is extended to 3262. In ACF, the shrinking parameter is $s = 4$. The Adaboost classifier with depth-2 decision tree trained in four rounds (32, 128, 512, 2048) is used. Since the pedestrian heights in Caltech dataset vary between 25 and 360 pixels, we should upscale by one octave to detect the pedestrians with fewer than 50 pixels. Since the pedestrian heights in TUD-Brussel dataset vary from 25 to 360 pixels, only the images that contain the pedestrians of heights larger than 50 pixels are upscaled by one octave for evaluation. In the implementation, the Piotr toolbox [50] is used to compute the channel feature maps and evaluate the detection methods, while the ResNet152 model is used as the refining network. The miss rate–FPPI curve is used to evaluate the performance of different methods. Note that the proposed methods based on Subset − Haar LFDA and ResNet are denoted by ACF + Subset − Haar, ACF + Subset − Haar + LFDA, and ACF + Subset − Haar + LFDA + ResNet, respectively.

Firstly, the threshold for IoU (Thr_pos and Thr_neg) is discussed. When ResNet is refined, the training samples are selected according to IoU. The samples with IoU $>$ Thr_pos are selected as the positive samples, while those with IoU $<$ Thr_neg are chosen as the negative samples. Thus, the two thresholds affect the quantity and quality of the training samples in ResNet training. If Thr_pos is larger and Thr_neg is smaller, the number of cropped images from region proposals will be smaller with higher confidence of training samples, and vice versa.

Afterwards, the influence of the padding operations on the obtained results is analyzed. Since some region proposals with IoU $>$ Thr_pos only contain partial areas of the pedestrian resulting in the lossy description and degrading the ResNet model performence, the padding operation is imposed on the region proposals with IoU $>$ Thr_pos such that the whole pedestrian can be cropped. However, if padding size is too large, the cropped proposals tend to contain more background areas, which also adversely affect the ResNet performance. In this section, ResNet152 is used in the implementation.

The ablation studies performed on the INRIA dataset are shown in Table 2. The obtained results can be summarized as follows. (1) With much tighter IOU between Gt and proposals, the best performance can be achieved at a setting of IOU-0.7-0.3. This indicates that a more accurate bounding box can eliminate false positives. In addition, the performance drops at the IOU-0.8-0.2 setting, which can also lead to false negatives due to the very strict IOU threshold. (2) The proper padding helps to improve the performance due to the context information involved in the cropped images. The larger padding size cannot further improve the performance in the higher IOU scenarios due to the relatively high resolution of training samples in the INRIA dataset. (3) The consistent results are found with padding size of 16 pixels along the images width and height, which can better improve the performance compared to the case where no padding exists.

Table 2. Ablation studies on the INRIA dataset.

Method	MR [10^{-2}, 10^{0}]
IOU-0.5-0.5	11.82
IOU-0.5-0.5-16px	10.45
IOU-0.5-0.5-32px	10.09
IOU-0.6-0.4	10.29
IOU-0.6-0.4-16px	9.22
IOU-0.6-0.4-32px	10.05
IOU-0.7-0.3	9.62
IOU-0.7-0.3-16px	8.6
IOU-0.7-0.3-32px	9.18
IOU-0.8-0.2	12.95
IOU-0.8-0.2-16px	11.31
IOU-0.8-0.2-32px	12.05

4.2. Discussion on Pedestrian Proposals

Table 3 presents a quantitative comparison of the pedestrian proposal approaches generated by ACF [51], HF, and WHF on the three benchmarks. It can be seen from Table 3 that HF and WHF outperform ACF in terms of the total number of proposals and recall rate on the INRIA, Caltech, and TUDBrussel datasets, where HF is short for Ours (subset-haar) and WHF is short for Ours (LFDA). In particular, WHF outperforms ACF by 2–4% in the recall score with 5–8 times lower proposal numbers, on average. Note that the recall is calculated with the height of the ground truth pedestrians having more than 50 pixels. In this paper, WHF is used as the shallow feature-based pedestrian proposal method.

Table 3. Pedestrian Proposals on Different Datasets.

Datasets	Images	Proposals			Recall			Average		
		ACF	HF	WHF	ACF	HF	WHF	ACF	HF	WHF
INRIA	288	60,243	6221	5138	95.59%	96.18%	96.72%	209.17	21.62	17.84
Caltech	4024	625,037	91,317	87,251	95.40%	96.54%	97.38%	155.32	22.69	21.68
Tud-Brussels	508	16,818	5133	3816	86.38%	87.82%	88.32%	33.11	10.12	7.51

4.3. Experimental Results

4.3.1. Results on INRIA Dataset

Figure 4 shows the detection performance of the proposed methods, Ours(ResNet152), Ours(Subset-Haar), purs(LFDA), and other classic algorithms from the literature (15 algorithms), on the INRIA dataset. It can be observed that the proposed methods outperform the other methods. More precisely, the miss rate of the Ours(ResNet152) method is the lowest (8.47%) among the 15 algorithms, and the FPPI is 0.1. Each stage of SDFF could further improve the performance. In addition, the average MR obtained by Ours(ResNet152) is lower by 5.96%, 5.32%, 5.23%, 5.06%, 4.85%, and 2.75% than that of InformedHaar [52], LDCF [18], Franken [53], Roerei [53], SketchToken [54], and Spatial Pooling [55], respectively. It can also be seen that Ours(ResNet152) has a higher performance than Ours(LFDA) method [7] with a 4.61% lower miss rate. This is due to the fact that the ResNet152 model performs secondary feature extraction and classification based on the Ours(LFDA) output. Therefore it enhances the performance of the classification.

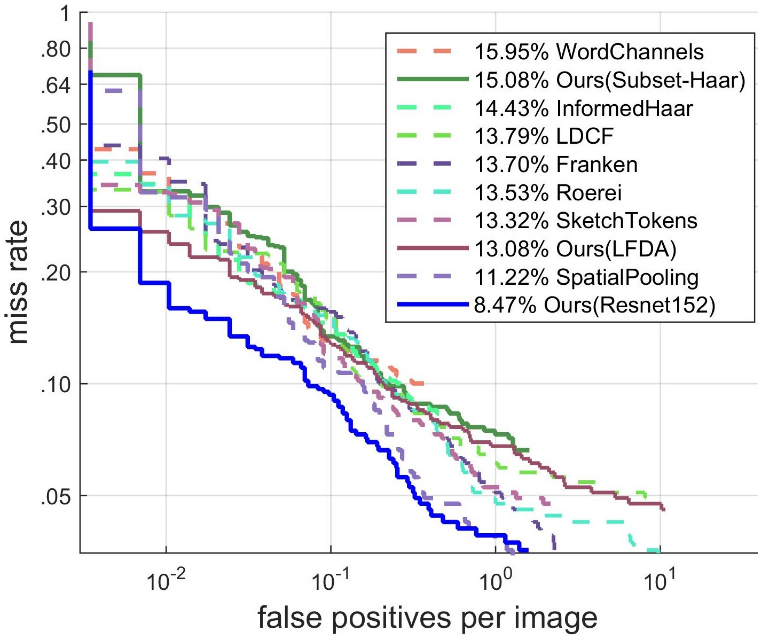

Figure 4. Comparison of different methods on the INRIA dataset.

4.3.2. Results on Caltech Dataset

The experiments on the Caltech dataset with the original annotations were conducted to evaluate the proposed methods. Similar to the results on the INRIA dataset, Ours(RestNet152) achieves the best results compared with the other methods, as shown in Figure 5. The corresponding performance advantages reach 12.1%, 7.31%, 7.05%, and 1.73% on average MR against Spatial Pooing [55], Checkerboards+ [17], MRFC + Semantic, and UDN+ [56], respectively. In addition, it can be seen that Ours(LFDA) is almost

5.13% higher than ours(RestNet152) in terms of average MR, which is consistent with the results achieved on the INRIA dataset. Moreover, the proposed method outperforms the deep models, e.g., CCF + CF [57] and TA-CNN [58]. Based on the observation, we can deduce that Ours(ResNet152) can efficiently prune the hard false positives.

The experiments on Caltech dataset with the new annotations [17] were also conducted for evaluation. As shown in Figure 5, the proposed method reduced the average miss rate 3.43% compared to the original annotations, which indicates the precision of the ground truth has a positive impact on the performance. More precisely, the proposed method is about 9.79%, 8.01%, and 7.52% lower in the miss rate than TA-CNN [58], MRFC + Semantic [22], and Checkerboards+ [17], respectively. In addition, the experimental results indicate that Ours(ResNet152) can effectively eliminate the false positives from the first cascade with high precision.

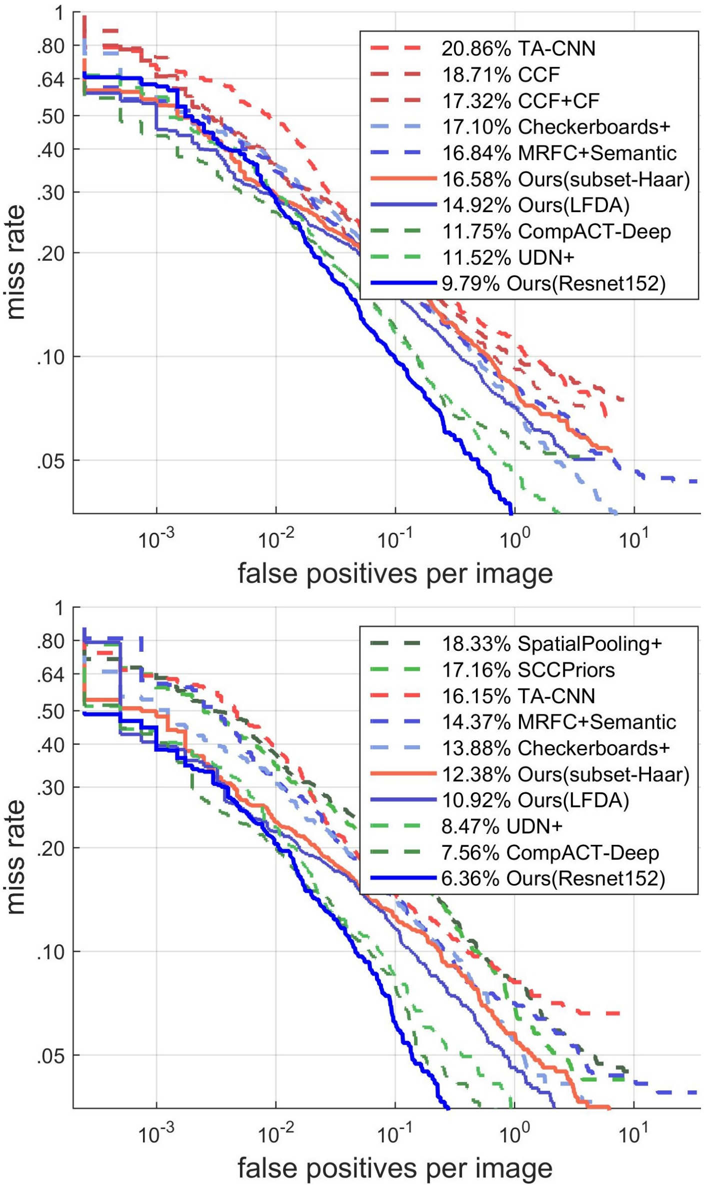

Figure 5. The performance comparison of different methods on the Caltech dataset. **Uppper**: original annotation, **Down**: new annotations.

4.3.3. Results on the TUD-Brussel dataset

The proposed methods were also compared with the classical methods on the TUD-Brussels dataset. These methods were trained on the INRIA dataset and tested using

508 images with 1326 annotated pedestrians. Due to the pedestrians' heights in the TUD-Brussel dataset within [25, 360] pixels, only the pedestrians whose heights are larger than 50 pixels were evaluated, and the corresponding images are upscaled by 1 octave in the testing stage. As shown in Figure 6, the proposed methods outperform ACF [16], LDCF [18], and MF + Motion + 2Ped [59] and result in comparable results with the Spatial Pooling [55] method. Figure 6 clearly shows that Ours(Restnet152) outperforms the Spatial Pooling method with approximately 0.92% lower average MR. Furthermore, the model trained on the INRIA dataset can be transferred to the TUD-Brussels dataset due to the complex variations included in this dataset.

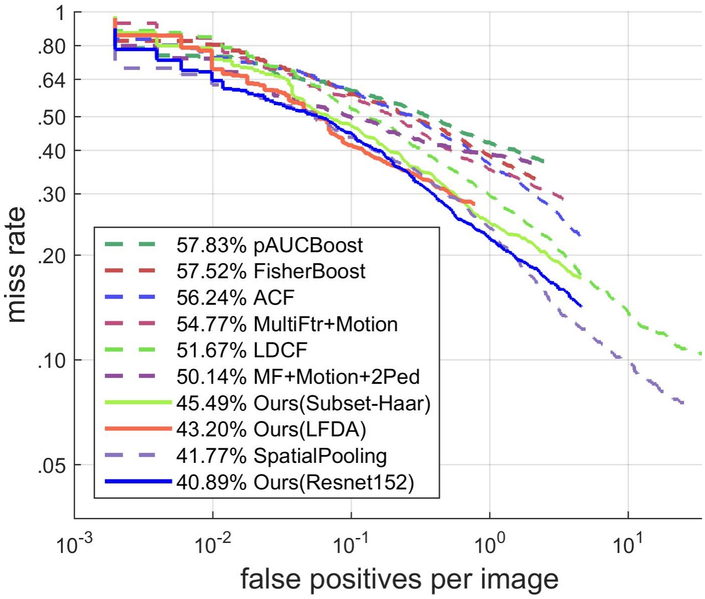

Figure 6. The performance comparison of different methods on the TUD-Brussel dataset.

4.3.4. Comparison with FRCNN, YOLOV4 and SSD

More experimental comparisons with the state-of-the-art methods are shown in Table 4. It can be seen that the proposed method outperforms the state-of-the-art methods on the INRIA, Caltech, and Tud-Brussel datasets.

Table 4. Comparison with state-of-the-art methods on Different Datasets.

	FRCNN + FPN	**YOLOV4**	**SSD512**	**Ours**
INRIA	11.82%	12.77%	19.25%	8.47%
Caltech	10.74%	11.61%	29.72%	6.36%
Tud-Brussel	48.70%	50.83%	54.44%	40.89%

Table 5 shows the speed on a GPU (Nvidia 1080 TI) of the above methods. It is clear to see that yolov4 is the fastest method, but our method is comparable to SSD512 and faster than the FRCNN + FPN method.

Table 5. Running speed Comparison of 640 × 480 images (P is short for proposal, R is short for refining).

Method	**FRCNN + FPN**	**YOLOV4**	**SSD512**	**Ours(P)**	**Ours(P + R)**
Speed (FPS)	15	46	22	32	21

5. Conclusions

Inspired by the combination of traditional pattern recognition and DCNN for image classification and object detection, a shallow–deep feature fusion method (SDFF) for pedestrian detection is proposed. The shallow feature, also referred to as the weighted subset-Gaar-like feature, is obtained by filtering the channel maps of ACF using Haar-like templates followed by Local FDA feature learning. Weighted subset-Haar-like features with ACF framework could be leveraged for efficient pedestrian detection. Furthermore, the shallow feature-based pedestrian detection method can be used as region proposal in object detection. ResNet is then used to refine the weighted subset-Haar-like features as a deep feature pedestrian detection method. Experiments were conducted on INRIA, Caltech, and TUD-Brussel datasets in order to to evaluate the proposed method. The experimental results show that each step could further enhance the performance of SDFF.

Author Contributions: Conceptualization, algorithms, methodology, writing—original draft preparation, D.L.; software, validation, K.Z.; writing-review & editing, D.L.; J.S.; supervision, J.S.; Funding acquisition, D.L.; J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partly supported by Science and Technology Development Fund of Macao S.A.R (FDCT) under number 0015/2019/AKP and NSF of Jiangsu Province in China under Grants BK20191427.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hasan, I.; Liao, S.; Li, J.; Akram, S.U.; Shao, L. Generalizable Pedestrian Detection: The Elephant in the Room. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 21–24 June 2021; pp. 11328–11337.
2. Lima, J.P.; Roberto, R.; Figueiredo, L.; Simoes, F.; Teichrieb, V. Generalizable Multi-Camera 3D Pedestrian Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Nashville, TN, USA, 21–24 June 2021; pp. 1232–1240.
3. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef] [PubMed]
4. Brunetti, A.; Buongiorno, D.; Trotta, G.F.; Bevilacqua, V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing* **2018**, *300*, 17–33. [CrossRef]
5. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
6. Felzenszwalb, P.F.; McAllester, D.A.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; Volume 2, p. 7.
7. Zang, K.; Shen, J.; Yang, W. Using LFDA to Learn Subset-Haar-Like Intermediate Feature Weights for Pedestrian Detection. In Proceedings of the International Conference on Intelligent Science and Big Data Engineering, Dalian, China, 22–23 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 215–230.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Oren, M.; Papageorgiou, C.; Sinha, P.; Osuna, E.; Poggio, T. Pedestrian detection using wavelet templates. In Proceedings of the Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; Volume 97, pp. 193–199.
10. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; pp. 511–518.
11. Zhao, K.; Deng, J.; Cheng, D. Real-time moving pedestrian detection using contour features. *Multimed. Tools Appl.* **2018**, *77*, 30891–30910. [CrossRef]
12. Zhu, Q.; Yeh, M.C.; Cheng, K.T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1491–1498.
13. Wang, X.; Han, T.X.; Yan, S. An HOG-LBP human detector with partial occlusion handling. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 32–39.
14. Bourdev, L.; Malik, J. Poselets: Body part detectors trained using 3d human pose annotations. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1365–1372.
15. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef] [PubMed]

16. Dollár, P.; Tu, Z.; Perona, P.; Belongie, S. Integral Channel Features. In Proceedings of the British Machine Conference, London, UK, 7–10 September 2009; pp. 91.1–91.11.
17. Zhang, S.; Benenson, R.; Schiele, B. Filtered channel features for pedestrian detection. *CVPR* **2015**, *1*, 4.
18. Nam, W.; Dollár, P.; Han, J.H. Local decorrelation for improved pedestrian detection. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014, pp. 424–432.
19. Shen, J.; Zuo, X.; Li, J.; Yang, W.; Ling, H. A novel pixel neighborhood differential statistic feature for pedestrian and face detection. *Pattern Recognit.* **2017**, *63*, 127–138. [CrossRef]
20. You, M.; Zhang, Y.; Shen, C.; Zhang, X. An extended filtered channel framework for pedestrian detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 1640–1651. [CrossRef]
21. Ouyang, W.; Zeng, X.; Wang, X. Modeling mutual visibility relationship in pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3222–3229.
22. Daniel Costea, A.; Nedevschi, S. Semantic channels for fast pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2360–2368.
23. Sheng, B.; Hu, Q.; Li, J.; Yang, W.; Zhang, B.; Sun, C. Filtered shallow-deep feature channels for pedestrian detection. *Neurocomputing* **2017**, *249*, 19–27. [CrossRef]
24. Park, D.; Zitnick, C.L.; Ramanan, D.; Dollár, P. Exploring weak stabilization for motion feature extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2882–2889.
25. Cheng, M.M.; Zhang, Z.; Lin, W.Y.; Torr, P. BING: Binarized normed gradients for objectness estimation at 300fps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3286–3293.
26. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 391–405.
27. Van de Sande, K.E.; Uijlings, J.R.; Gevers, T.; Smeulders, A.W. Segmentation as selective search for object recognition. *ICCV* **2011**, *1*, 7.
28. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2189–2202. [CrossRef] [PubMed]
29. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1312–1328. [CrossRef] [PubMed]
30. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
31. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
32. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012; pp. 1097–1105.
34. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
35. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
36. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *arXiv* **2018**, arXiv:1809.02165.
37. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
38. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
39. Cheng, B.; Wei, Y.; Shi, H.; Feris, R.S.; Xiong, J.; Huang, T.S. Revisiting RCNN: On Awakening the Classification Power of Faster RCNN. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
40. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster r-cnn doing well for pedestrian detection? In Proceedings of the European Conference on Computer Visio, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 443–457.
41. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 354–370.
42. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
43. Sugiyama, M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.* **2007**, *8*, 1027–1061.
44. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 711–720. [CrossRef]

45. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Training very deep networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2377–2385.
46. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. How far are we from solving pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Amsterdam, The Netherlands, 11–14 October 2016; pp. 1259–1267.
47. INRIA Person Dataset. Available online: http://pascal.inrialpes.fr/data/human/ (accessed on 30 September 2021).
48. Caltech Pedestrian Detection Benchmark. Available online: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/ (accessed on 30 September 2021).
49. Wojek, C.; Walk, S.; Schiele, B. Multi-cue onboard pedestrian detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 794–801.
50. Piotr's Computer Vision Matlab Toolbox. Available online: http://pdollar.github.io/toolbox/ (accessed on 30 September 2021).
51. Dollar, P.; Appel, R.; Belongie, S.; Perona, P. Fast Feature Pyramids for Object Detection. *IEEE TPAMI* **2014**, *36*, 1532–1545. [CrossRef] [PubMed]
52. Zhang, S.; Bauckhage, C.; Cremers, A.B. Informed Haar-Like Features Improve Pedestrian Detection. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 947–954.
53. Mathias, M.; Benenson, R.; Timofte, R.; Gool, L.V. Handling Occlusions with Franken-Classifiers. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1505–1512.
54. Lim, J.J.; Zitnick, C.L.; Dollár, P. Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Sydney, Australia, 1–8 December 2013; pp. 3158–3165.
55. Paisitkriangkrai, S.; Shen, C.; van den Hengel, A. Pedestrian Detection with Spatially Pooled Features and Structured Ensemble Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1243–1257. [CrossRef] [PubMed]
56. Ouyang, W.; Zhou, H.; Li, H.; Li, Q.; Yan, J.; Wang, X. Jointly Learning Deep Features, Deformable Parts, Occlusion and Classification for Pedestrian Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1874–1887. [CrossRef] [PubMed]
57. Yang, B.; Yan, J.; Lei, Z.; Li, S.Z. Convolutional Channel Features. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 82–90.
58. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Pedestrian detection aided by deep learning semantic tasks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Santiago, Chile, 7–13 December 2015; pp. 5079–5087.
59. Walk, S.; Majer, N.; Schindler, K.; Schiele, B. New features and insights for pedestrian detection. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1030–1037.